EDGEWOOD
RESEARCH, DEVELOPMENT & ENGINEERING CENTER
U.S. ARMY CHEMICAL AND BIOLOGICAL DEFENSE COMMAND

ERDEC-CR-135

CONVEX-CONE CLASSIFICATION OF PYROLYSIS MASS SPECTRA OF BIOLOGICAL AGENTS

Michael L. Mavrovouniotis

NORTHWESTERN UNIVERSITY
Evanston, IL 60208-3120

Alice M. Harper
Agustin I. Ifarraguerri

RESEARCH AND TECHNOLOGY DIRECTORATE

August 1994

Approved for public release; distribution is unlimited.

CUM SCIENTIA DEFENDIMUS

Aberdeen Proving Ground, MD 21010-5423

Disclaimer

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorizing documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
1994 August

3. REPORT TYPE AND DATES COVERED
Final, 93 Jun - 93 Aug

4. TITLE AND SUBTITLE
Convex-Cone Classification of Pyrolysis Mass Spectra of Biological Agents

5. FUNDING NUMBERS
C-DAAL03-91-C-0034
TCN 93-132

6. AUTHOR(S)
Mavrovouniotis, Michael L. (Northwestern University);
Harper, Alice M.; and Ifarraguerri, Agustin I. (ERDEC)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Northwestern University, 2145 Sheridan Road,
Evanston, IL 60208-3120

DIR, ERDEC, ATTN: SCBRD-RTM, APG, MD 21010-5423

8. PERFORMING ORGANIZATION REPORT NUMBER
ERDEC-CR-135

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

DIR, ARO, P.O. Box 12211, Research Triangle Park, NC 27709

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

COR: Dr. Alice M. Harper, SCBRD-RTM, (410) 671-4115

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This work addressed the classification of biological samples using high-dimensional, time-dependent pyrolysis mass spectra. The data were projected onto a low-dimensional subspace using singular value decomposition. Then, a convex cone was formed on this subspace showing, as its corners, physically meaningful components of the sample. This technique enabled separation of a biological-material signal, largely independent of the absolute amount of sample. The detection of the presence of any biological material could be accomplished based on the convex cone alone, without other reference to the mass spectra. Automated clustering of samples was successfully carried out using a minimal spanning tree.

14. SUBJECT TERMS

Pyrolysis	Subspaces	Biological material identification
Convexity	Mass spectroscopy	
Cones	Multivariate analysis	

15. NUMBER OF PAGES
50

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UL

Blank

PREFACE

The work described in this report was authorized under Contract No. DAAL03-91-C-0034, TCN 93-132. This work was started in June 1993 and completed in August 1993.

The use of trade names or manufacturers' names in this report does not constitute an official endorsement of any commercial products. This report may not be cited for purposes of advertisement.

This report has been approved for release to the public. Registered users should request additional copies from the Defense Technical Information Center; unregistered users should direct such requests to the National Technical Information Service.

Acknowledgments

This work was supported by the U.S. Army Edgewood Research, Development and Engineering Center (Dr. Alice M. Harper) under the auspices of the U.S. Army Research Office Scientific Services Program administered by Battelle.

Blank

CONTENTS

LIST OF FIGURES AND TABLES

Figures

Tables

CONVEX-CONE CLASSIFICATION OF PYROLYSIS MASS SPECTRA OF BIOLOGICAL AGENTS

1. INTRODUCTION

A difficulty arising in the detection of biological warfare agents is that biological samples contain in high proportion biological macromolecules, mainly in the form of protein, RNA, and lipids (Lehninger, 1982). For example, *E. Coli*, which is the most thoroughly studied bacterium, contains 96.1% (on a dry weight basis) macromolecules, 2.9% small organic molecules, and 1.0% inorganic ions (Ingraham *et al.*, 1983). The high molecular weight makes it difficult to analyze biological samples with techniques that require evaporation of the sample. A viable option is the use of pyrolysis in lieu of evaporation, in tandem with mass spectroscopy (Voorhees *et al.*, 1992). This report examines the analysis and classification of such spectra of biological materials. It expands on a basic approach described in an earlier report by Mavrovouniotis et al(1993).

The measurements considered here were obtained by the U.S. Army's Chemical and Biological Mass Spectrometer or CBMS (Sickenberger *et al.*, 1992), in the form of a continuous time profile of mass spectra, for successive pyrolysis cycles. A mass spectrum usually shows intensities for fragments grouped by m/Z (mass/charge) ratios. In the case of the CBMS, ions with unit charge are dominant, and we may therefore refer to "masses" rather than "mass/charge ratios" in the spectrum.

It is not clear, at the outset whether having measurements in the form of a time sequence of mass spectra (rather than a single spectrum) is an advantage or disadvantage. Given that the size of the sample might vary, but various interferences do not vary in proportion to the sample, it may be possible to extract information that is descriptive of the sample character without undue influence by its quantity. This would correspond to deconvoluting the signal most characteristic of the biological sample from background signals. We note here that background includes not only other particulates but, more importantly, mass spectra that are generated by components of the device itself as well as remnants of previous samples.

Ultimately, one would like to extract specific pyrolysis components, and identify the biochemical molecules and macromolecules of the sample. However, the complexity of a Py-MS signal and the fact that any biological sample is a complex mixture, makes this a difficult task. A first step in this direction would require a detailed model of pyrolysis of biochemical compounds.

This work addressed the analysis of high-dimensional time-dependent pyrolysis mass spectra of biological samples. We will see that the usual projection of the data onto a low-dimensional subspace using principal components analysis or singular value decomposition was followed by the construction of a convex cone which contains only physically meaningful spectra. This technique enabled accurate and robust separation of the signal most characteristic of the biological material.

The MATLAB package was used throughout this work. All of the techniques developed for this project were implemented as MATLAB scripts or function (M-files) and can be found in the Appendix, or in a previous report (Harper and Mavrovouniotis, 1993). Instead of using specialized routines for Principal Components Analysis from the Chemometrics Toolbox, we opted to use Singular Value Decomposition (SVD) which does much the same thing. SVD (Strang, 1988) decomposes any complex matrix A (which is allowed to be rectangular and/or singular) of dimensions m×n into the product:

$$A=U\Sigma V^T$$

where U (m×m) and V (n×n) are orthogonal matrices and Σ (m×n) is a matrix with zero off-diagonal elements. The elements of the main diagonal of Σ are non-negative, and they are called the singular values of A. If r is the rank of A, then only the first r singular values are non-zero.

For chemometric purposes, SVD effectively carries out Principal Components Analysis (PCA) (Jolliffe, 1986). If we view the columns of A to be samples, the columns of U define the eigenvectors (principal components) of the data, while each column i of $V\Sigma$ provides the projection of all the samples onto that eigenvector (component) i. Here, we will use the terms PCA and SVD interchangeably; we will also use the term "factor" as synonymous with the term "principal component".

2. FULL-INTENSITY SAMPLE

The data are originally in the form of MATLAB files, each describing a contiguous period of operation of the instrument, covering several pyrolysis cycles; Figure 1 shows the total intensity profile from one such file, identified as t15100. A first step that must be taken is the organization of the data into distinct pyrolysis cycles (Harper and Mavrovouniotis, 1993), which are identified by concatenating (at the end of the original symbol) the letter "p" and a two-digit number (taking the values 01, 02, etc.). For t15100 (Figure 1) the pyrolysis cycles are identified as t15100p01, t15100p02, t15100p03, t15100p04, t15100p05, t15100p06, t15100p07, t15100p08. Among the pyrolysis cycles in any given initial file, the first one or two usually precede the presence of the actual sample (in this case *MS-2 Coliphage*), i.e., they are background cycles. Some cycles after those might contain a partial amount of sample. Figure 1 shows that for t15100 the first two cycles (t15100p01, t15100p02) are background, the next one (t15100p03) is partial sample, and the remaining five complete cycles (t15100p04-t15100p08) are full sample. The pre-processing routines additionally trim the spectrum to the mass range 46-149, which contains the most significant information.

We analyze here the last complete pyrolysis cycle, i.e., t15100p08. This section is a summary of the analysis carried out by Harper and Mavrovouniotis (1993). A mesh plot for sample t15100p08 is shown in Figure 2. It shows that after a brief initial period of high intensity (much higher for some masses than for others) all the intensities decay.

The approach for isolating a signal characteristic of the sample is based on projection onto a subspace followed by construction of a convex cone (Harper and Mavrovouniotis, 1993). We first scale the spectrum of each time-point, to make its total ion count equal to one. An effect of the normalization is that our study is from now on independent of the absolute level of the Py-MS signal; any background deconvolution, detection, and classification will have to depend on the composition and not the amount of the ion stream. Then, we carry out SVD of the time-evolving spectrum; the 18 largest singular values for the sample t15100p08 are shown in Table 1. The first 3 factors explain 99.0% of the variance, and the first 18 (out of a total of 100) explain 99.8% of the variance. The time profiles for first three principal components of t15100p08 are shown in Figure 3. The correlation in this and subsequent time-profiles can simply be viewed as indicating the angle formed by the two vectors, with a value of 1 indicative of perfect alignment.

In our approach, we construct not only a subspace by taking the first few factors but in fact a convex cone whose interior corresponds precisely to the set of physically meaningful spectra on the subspace. Letting the spectra of the first three factors, u_1, u_2, and u_3, we can form $u_1+gu_2+hu_3$. We know that u_1 is positive; the extreme values of g and h define the corners of a polygon in two dimensions. The cone of acceptable spectra for t15100p08 is shown in precisely this form, i.e., in terms of coefficients attached to Factor 2 and Factor 3, by Figure 4. In effect, we are showing the intersection of a three-dimensional cone with the plane in which Factor 1 is equal to 1. Legitimate spectra lie in the interior of the convex irregular pentagon of Figure 4.

The time-profiles of all the corners are shown in Figure 5. Figures 6 to 10 show, in succession, the spectra and time-profiles of the 5 extreme points, in the order defined by Figure 4.

Text continues on page 19.

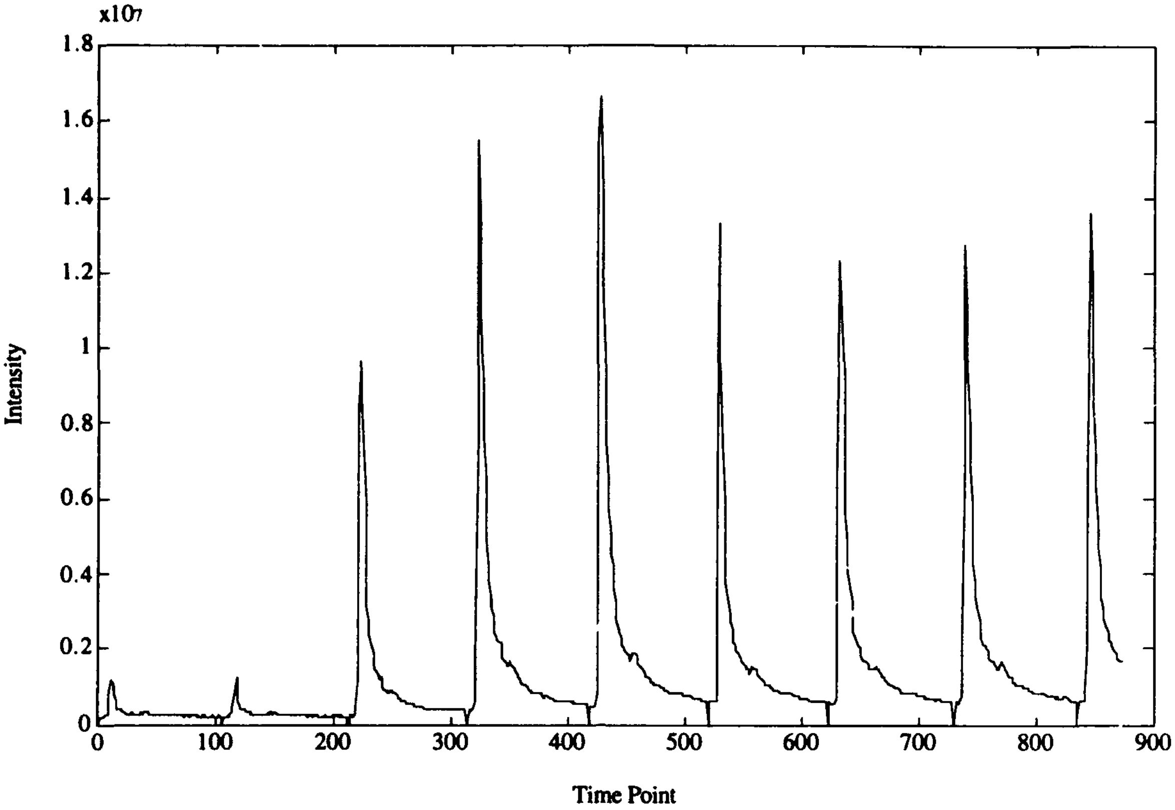


Figure 1. Total intensity for the sample series t15100.

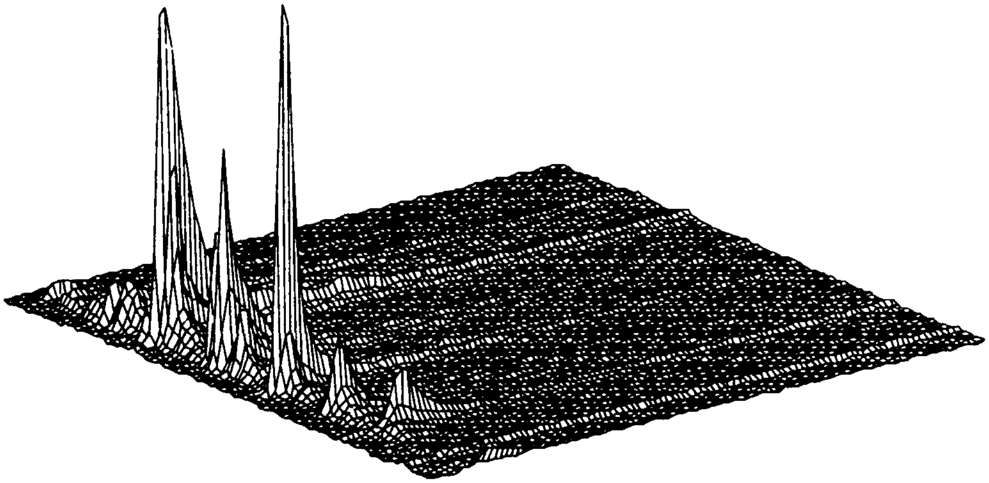


Figure 2. The spectr'm for file t15100p08. The time axis is oriented from the front corner towards the right. The mass axis is in the front left.

Table 1. The 18 largest singular values for the normalized sample t15100p08.

σ_1 to σ_9	σ_{10} to σ_{18}
9.7564	0.1574
1.6791	0.1468
0.9936	0.1425
0.5668	0.1325
0.3925	0.1303
0.2660	0.1232
0.1926	0.1154
0.1831	0.1150
0.1625	0.1137

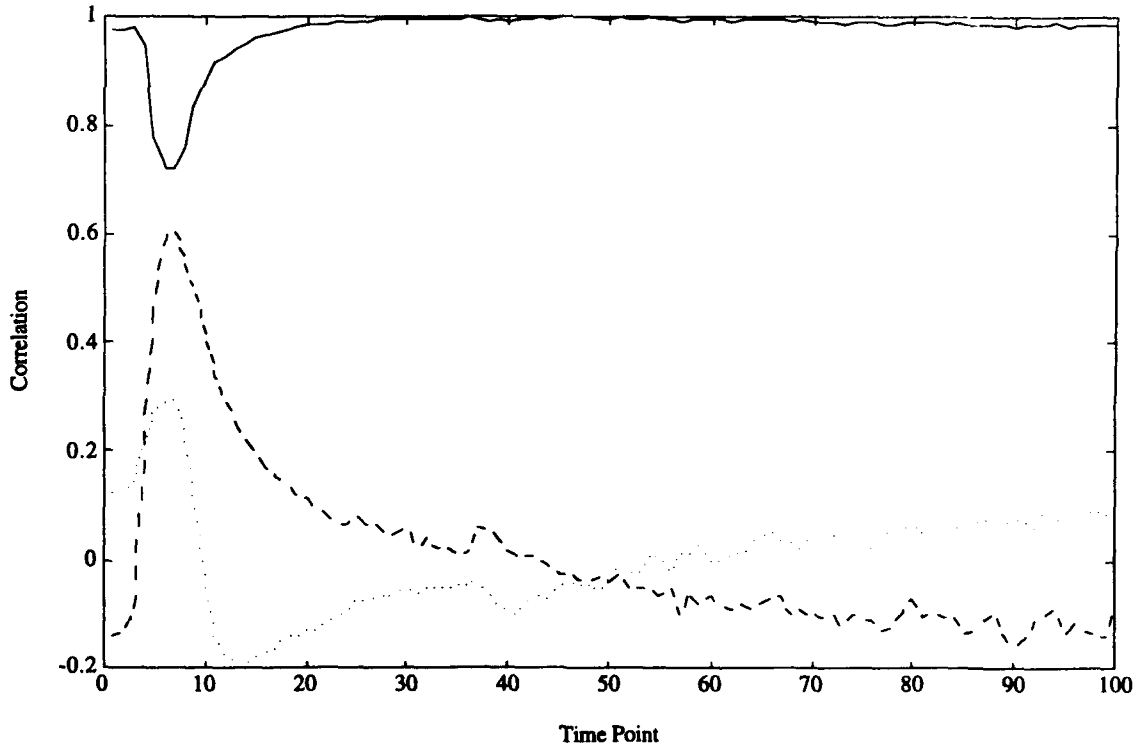


Figure 3. *Time profiles for the first three principal components of t15100p08*

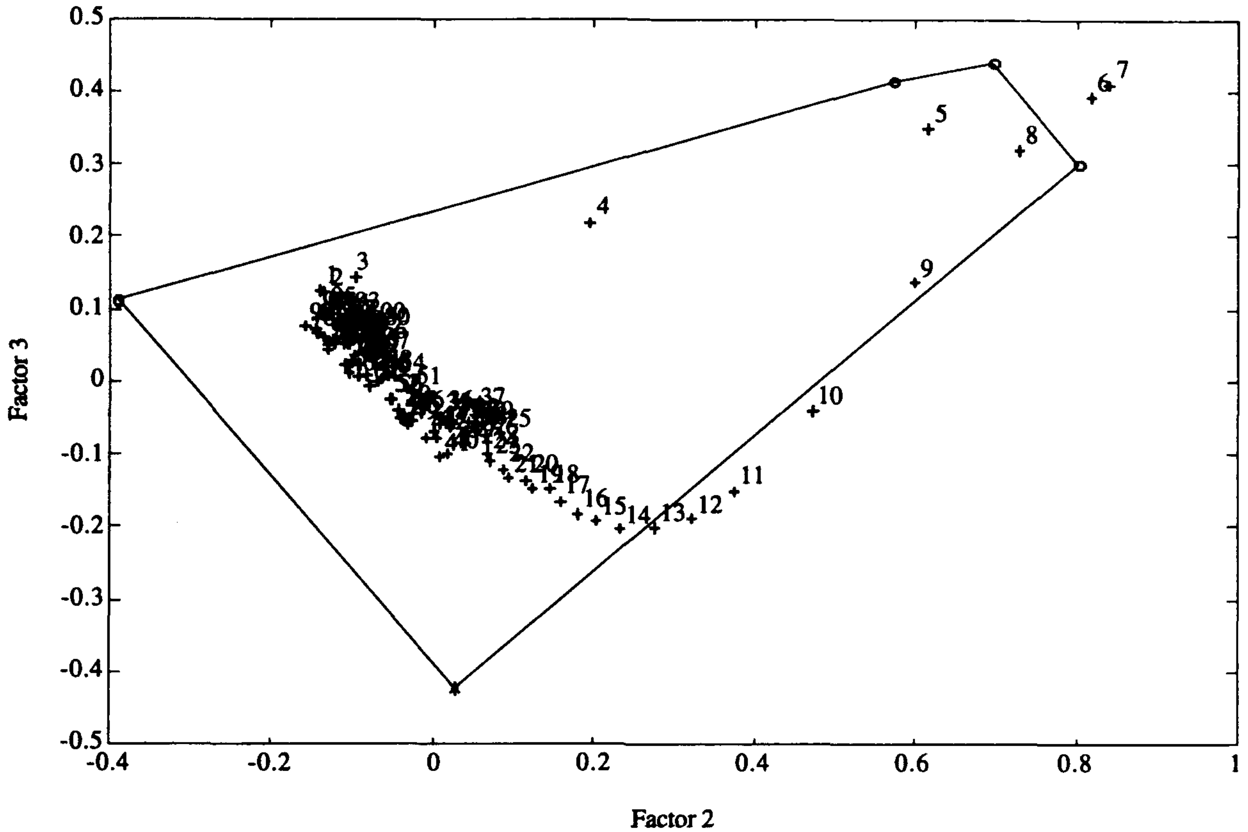


Figure 4. The cone of acceptable spectra for t15100p08 using three factors. It is shown here in the form of the coefficients of Factor 2 and Factor 3, while the coefficient of Factor 1 is preset to 1. The first corner of the convex irregular pentagon is marked by a star and corners 2-5 (counterclockwise) are identified by circles. The projections of the actual spectra are also shown.

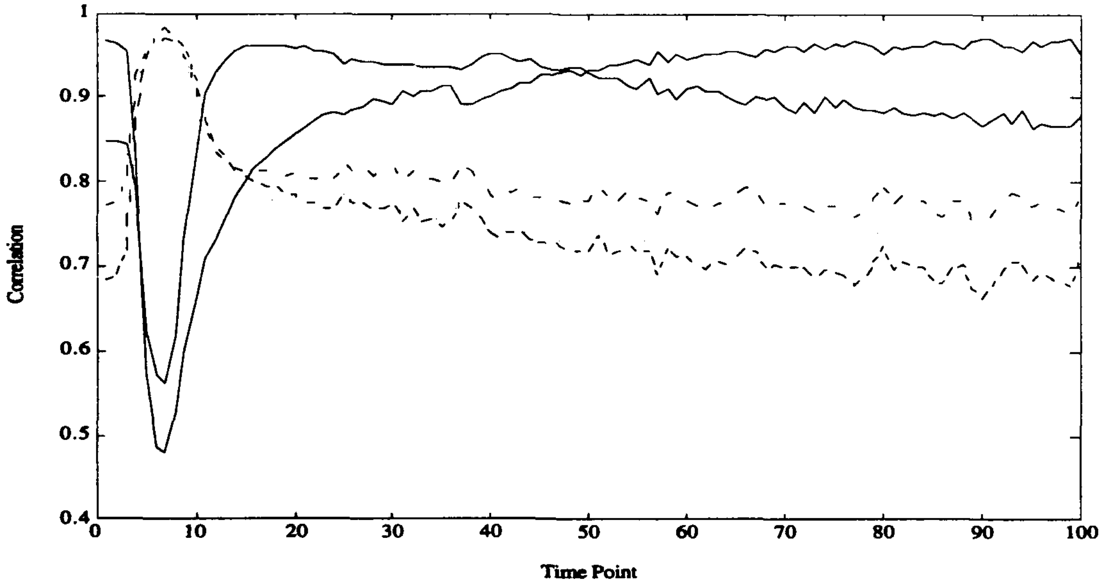


Figure 5. The correlation time-profiles of all the corners of the three-dimensional convex cone (defined by 5 extreme spectra) for sample t15100p08. These are shown in Figures 15 to 19.

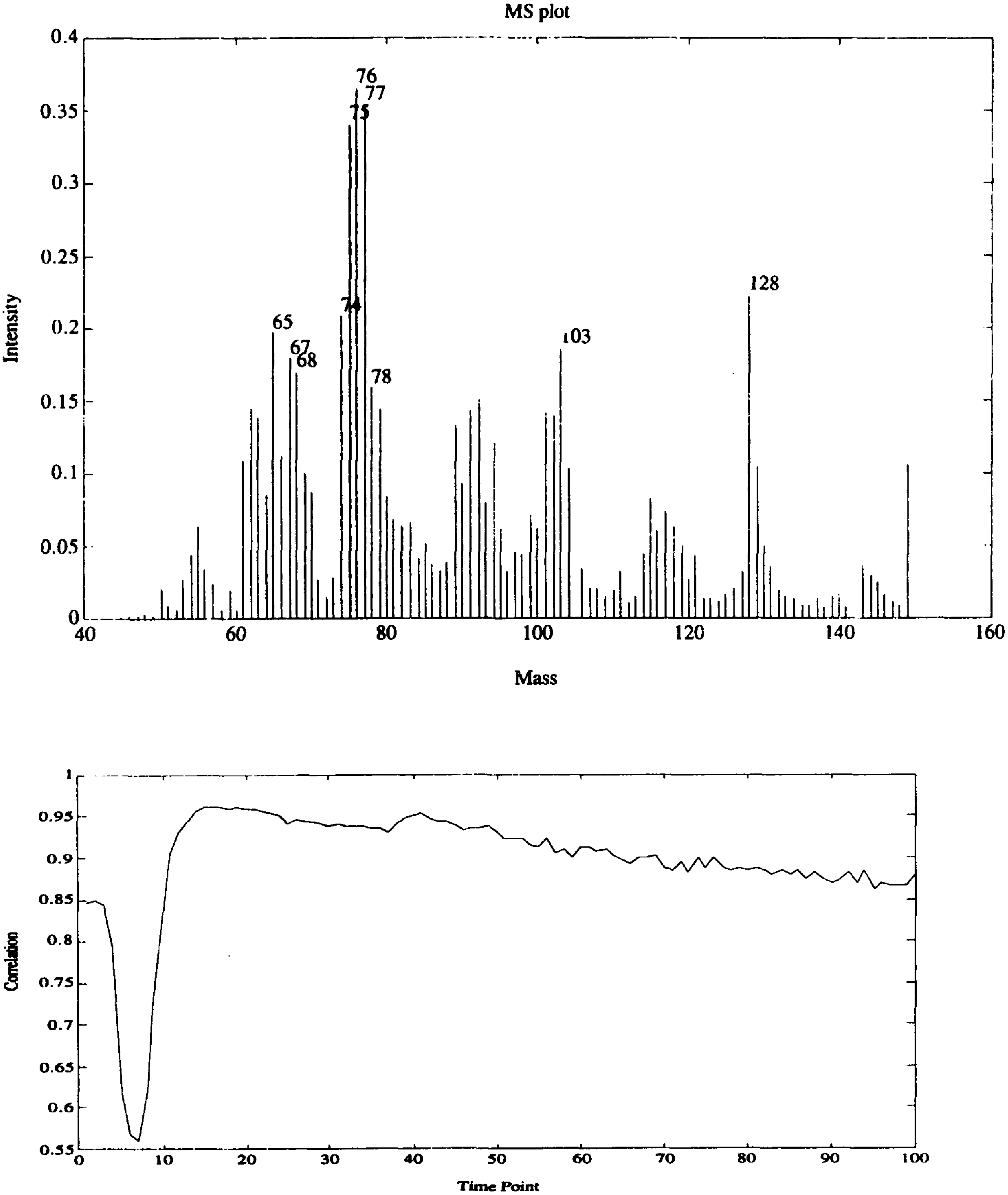
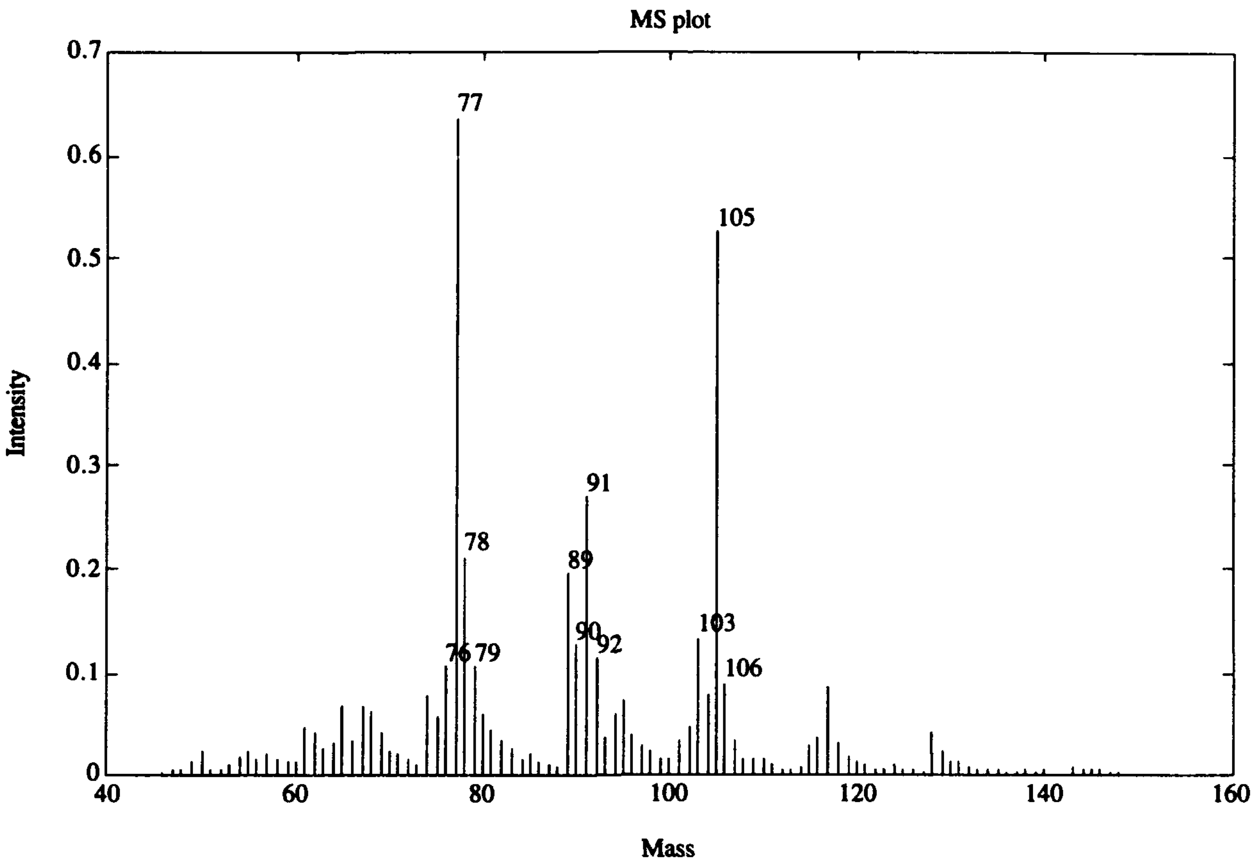


Figure 6. The spectrum and time-profile of the first extreme point of the three-dimensional convex cone of spectra (for sample t15100p08). This is the corner identified by a star near the bottom of Figure 4.



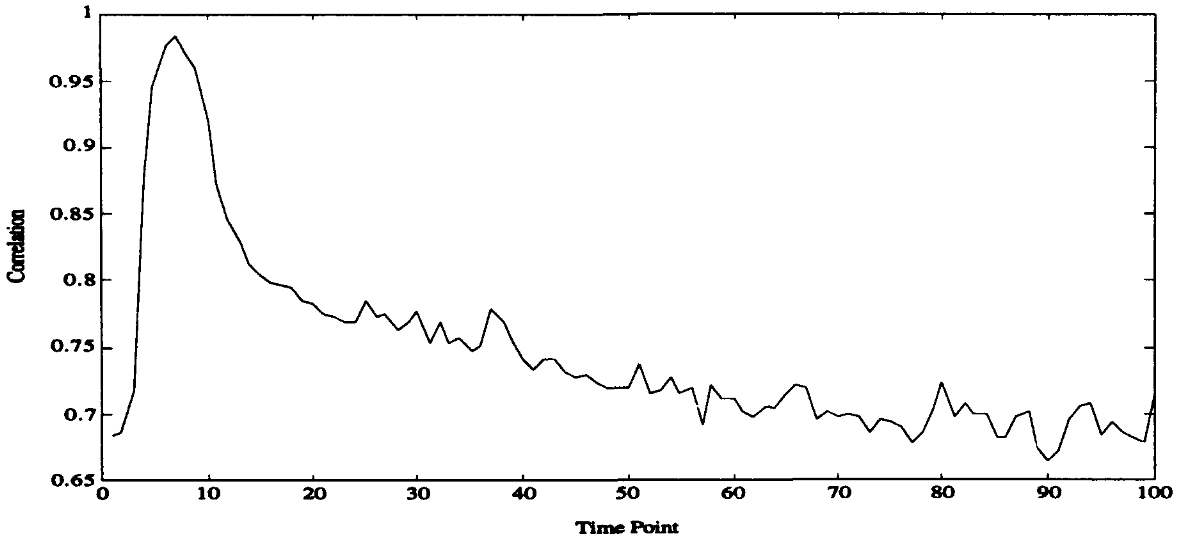


Figure 7. The spectrum and time-profile of the second extreme point of the three-dimensional convex cone of spectra (for sample t15100p08). This is the rightmost corner identified by a circle in Figure 4.

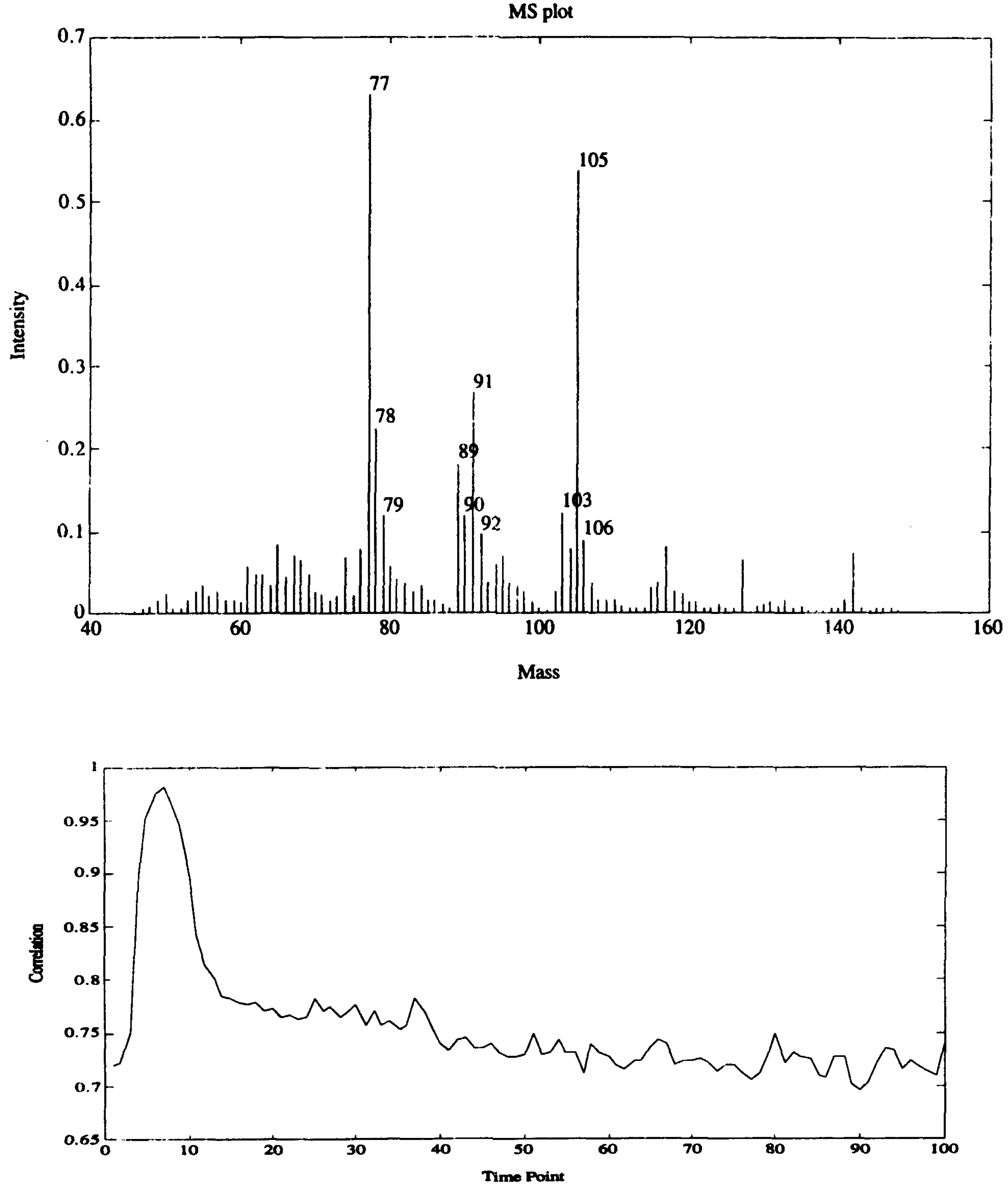


Figure 8. The spectrum and time-profile of the third extreme point of the three-dimensional convex cone of spectra (for sample t15100p08). This is the topmost corner identified by a circle in Figure 4.

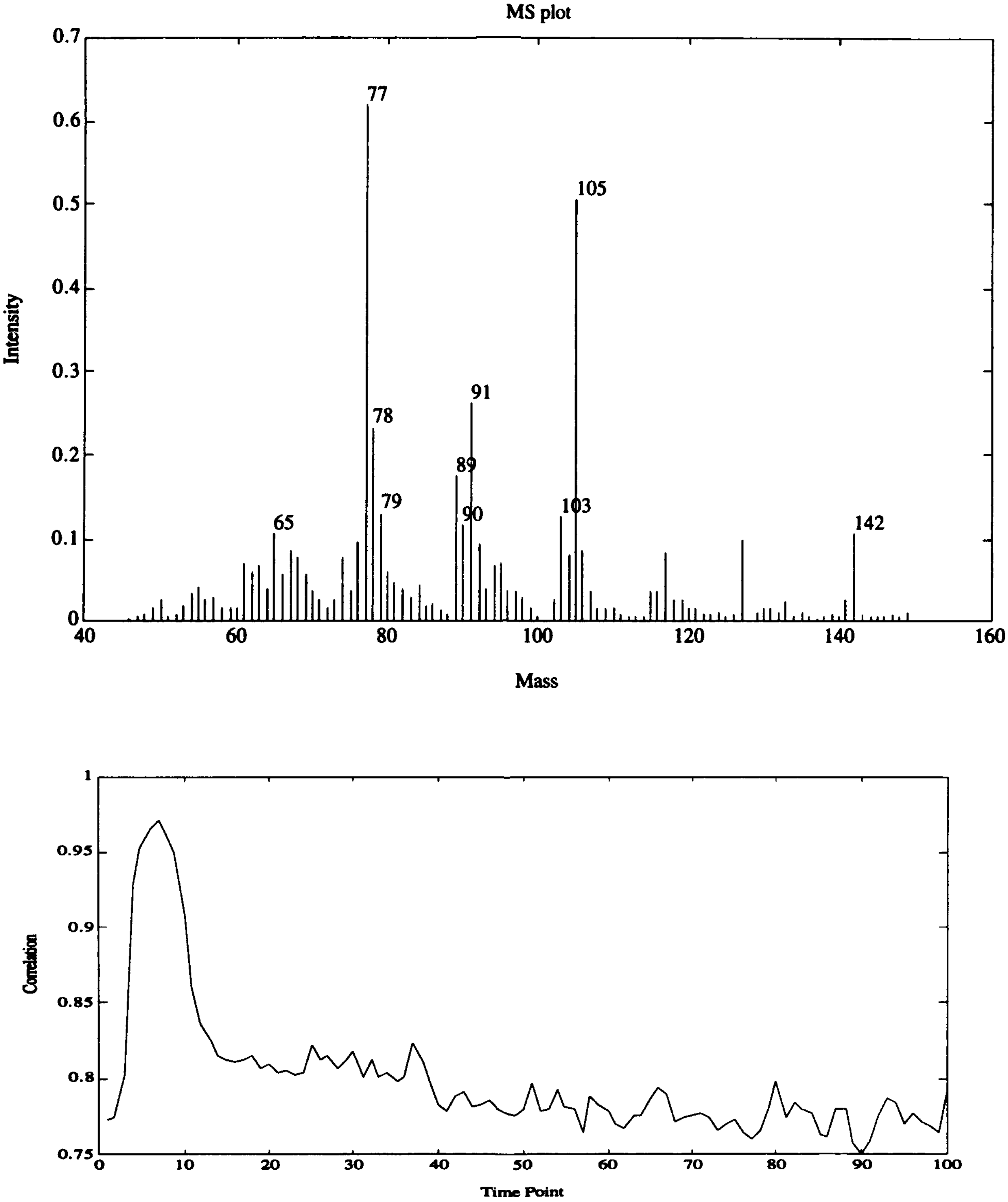


Figure 9. The spectrum and time-profile of the fourth extreme point of the three-dimensional convex cone of spectra (for sample t15100p08). This is the corner that forms a very obtuse angle between the topmost and leftmost corners (all corners identified by circles in Figure 4).

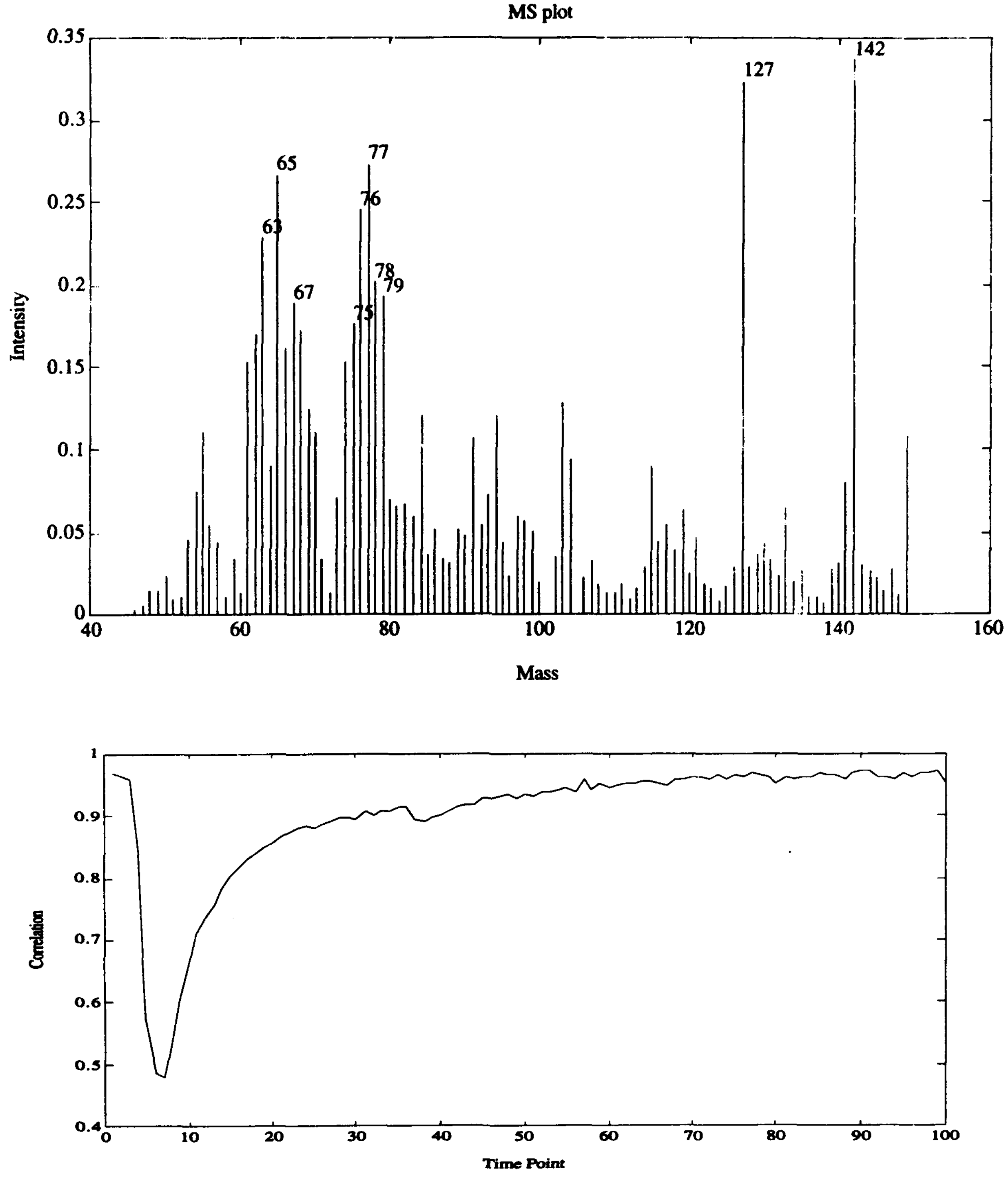


Figure 10. The spectrum and time-profile of the fifth extreme point of the three-dimensional convex cone of spectra (for sample t15100p08). This is the leftmost corner identified by a circle in Figure 4.

In Figure 4, it is clear that corners 2, 3, and 4 (circles in the upper right region of the pentagon) are very close to each other. If we select just one of these three then, in conjunction with corners 1 and 5, we will form a triangle which includes most of the area of the original polygon. If we take the location, shape, and size of the polygon as representative of its information content, the triangle will match well the information in the polygon. We find that this is generally the case with datasets that contain a biological-material sample (regardless of the nature of the material): The shape we obtain in the Factor 2 / Factor 3 space is very similar to a triangle. The same idea is reflected, in Figure 5, by the fact that we have three kinds of time profiles. The first displays a minimum in the region of highest intensity (time point 8), then a maximum shortly afterwards (time points 15-20), followed by steady gradual decline. The second displays a maximum in the region of time point 8 with a decline thereafter (there are three profiles with this behavior). The third has a minimum around time point 8 and then rises gradually. This clustering of the time profiles of Figure 5 corresponds precisely to the clustering of the corner points in Figure 4. The clustering of the corner points into three categories is, finally, evident in their spectra, in Figures 6 to 10.

Profiles that display a maximum in the region of time point 8 with a decline thereafter (the second category mentioned above) coincide with the peak total ion intensity. They model primarily the background that is caused by remnants of past samples as well as portions of the device itself. The reason is that as the temperature is elevated and the pyrolysis cycle begins, a number of phenomena will take place faster than the decomposition of the main biological sample. These phenomena include desorption of compounds adsorbed in various parts of the device as well as pyrolysis of non-biological particulates or non-particulates. These phenomena will thus account for the earliest portion of the rise in the ion intensity.

Profiles that display a minimum in the region of highest intensity (time point 8), then a maximum shortly afterwards (time points 15-20), followed by steady gradual decline, are characteristic of the actual biological material sample (they are the second category mentioned in the previous subsection). The delay in the pyrolysis (relative to faster, physical phenomena) allows these signals to be separated from the background. Profiles that have a minimum around time point 8 and then rise gradually are essentially the echo that compensates for the behavior of the first two types.

3. PARTIAL-INTENSITY SAMPLE

We stated earlier that the convex-cone methodology we followed for the derivation of a spectrum most characteristic of the biological material is robust with respect to the amount of sample. We will now demonstrate the validity of this claim, by examining sample t15100p03.

Referring to Figure 1, we can see that sample t15100p03 (the third peak in the figure) contains some of the biological material but has not reached the full-level intensity of subsequent samples; its peak intensity is only ~60% of the next peak, t15100p04. Our sample has only 95 time-points, and its mesh plot is shown in Figure 11.

3.1. Principal Components of Partial-Intensity Sample

The singular values (Table 2) show an elevation of the importance of factors 2 and 3 (compared to Table 1). The change in the behavior of these factors is also evident in the altered shapes of their time profiles (Figure 12 compared to Figure 3).

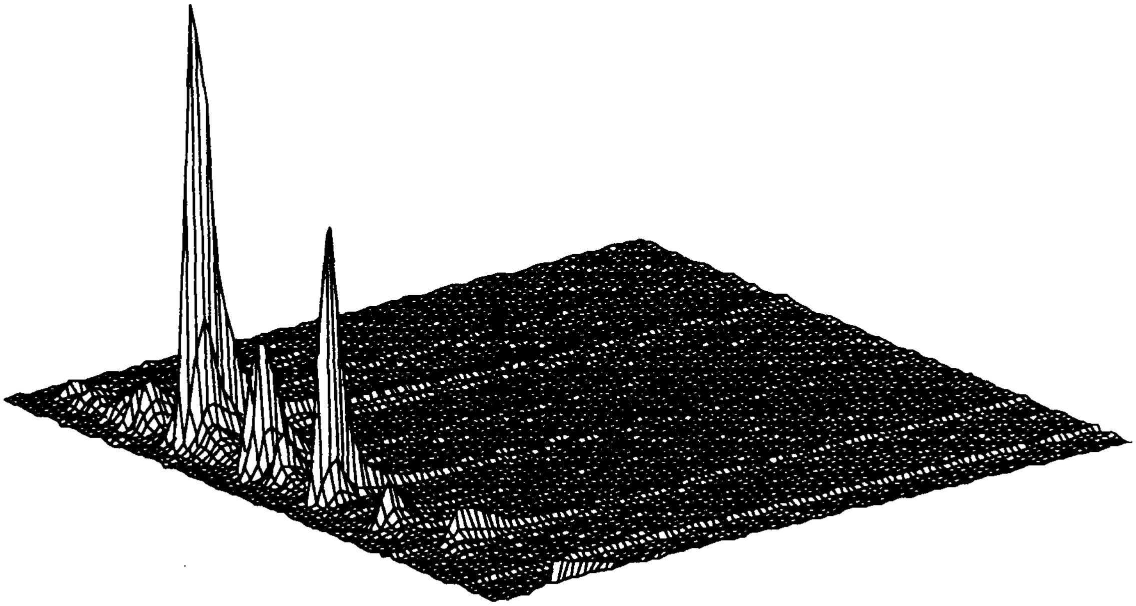


Figure 11. The spectrum for file t15100p03. The time axis is oriented from the front corner towards the right. The mass axis is in the front left.

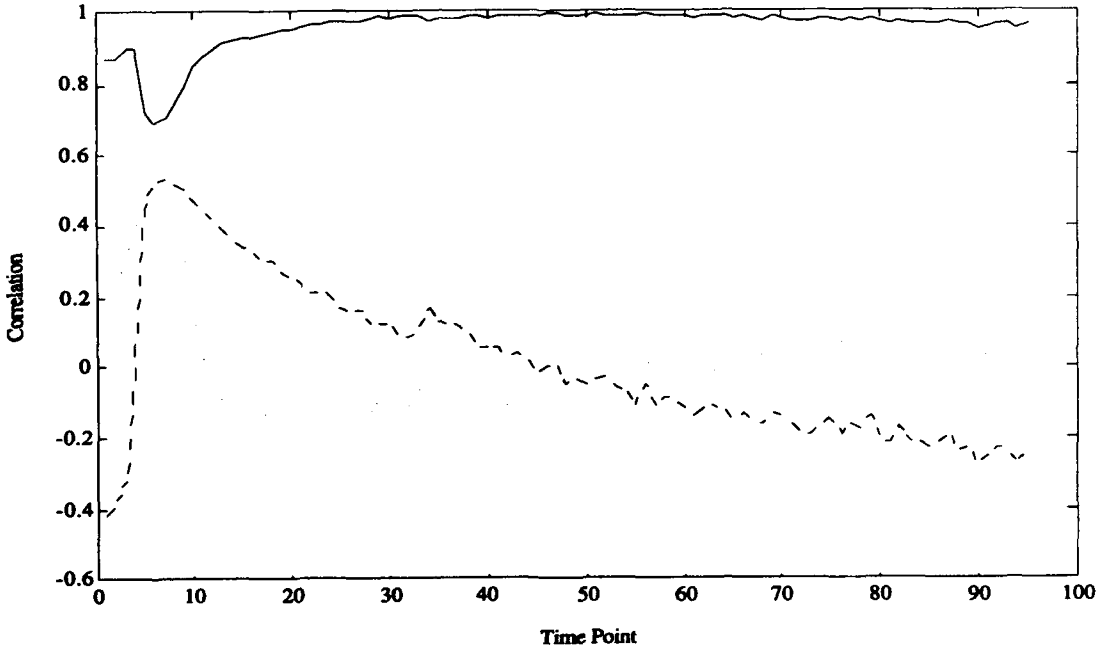


Figure 12. Time profiles for the first three principal components of t15100p03.

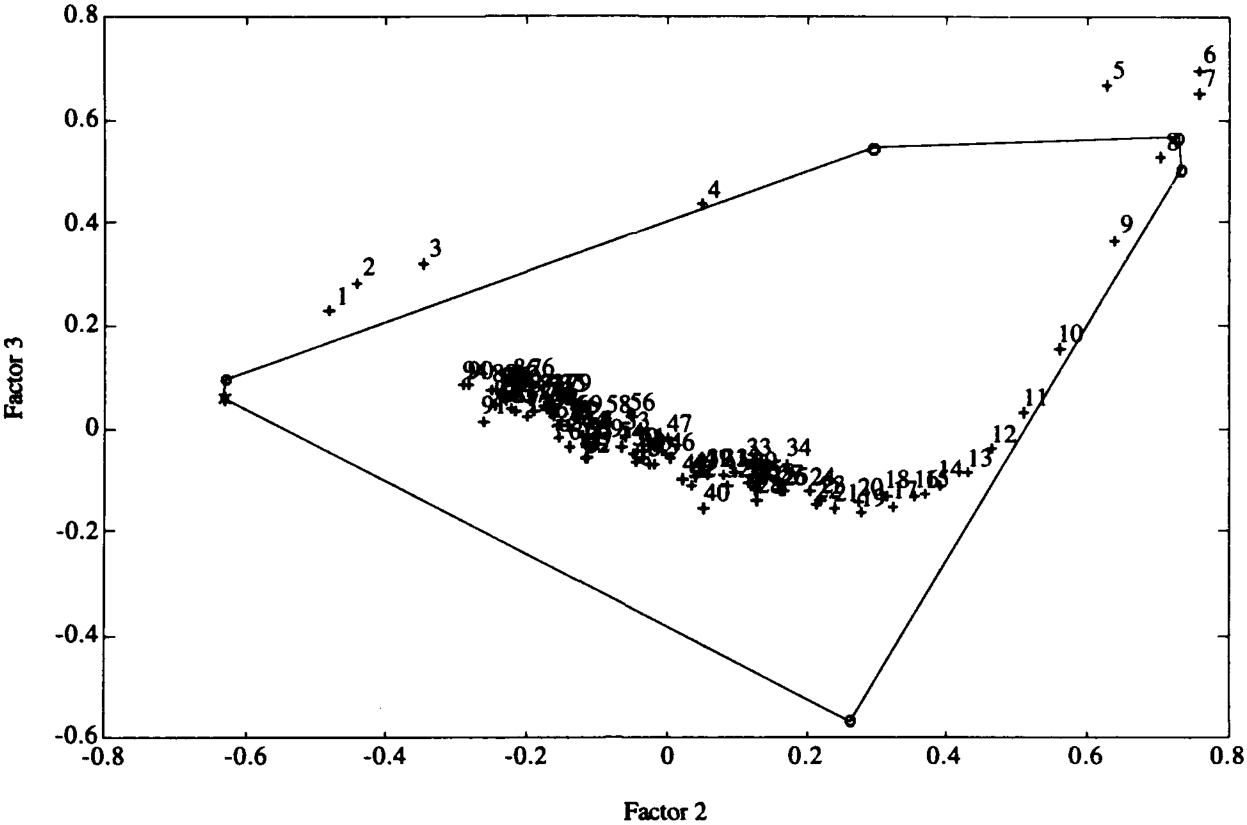


Figure 13. The three-dimensional cone of spectra for t15100p03 (shown in the form of the coefficients of Factor 2 and Factor 3, while the coefficient of Factor 1 is preset to 1).

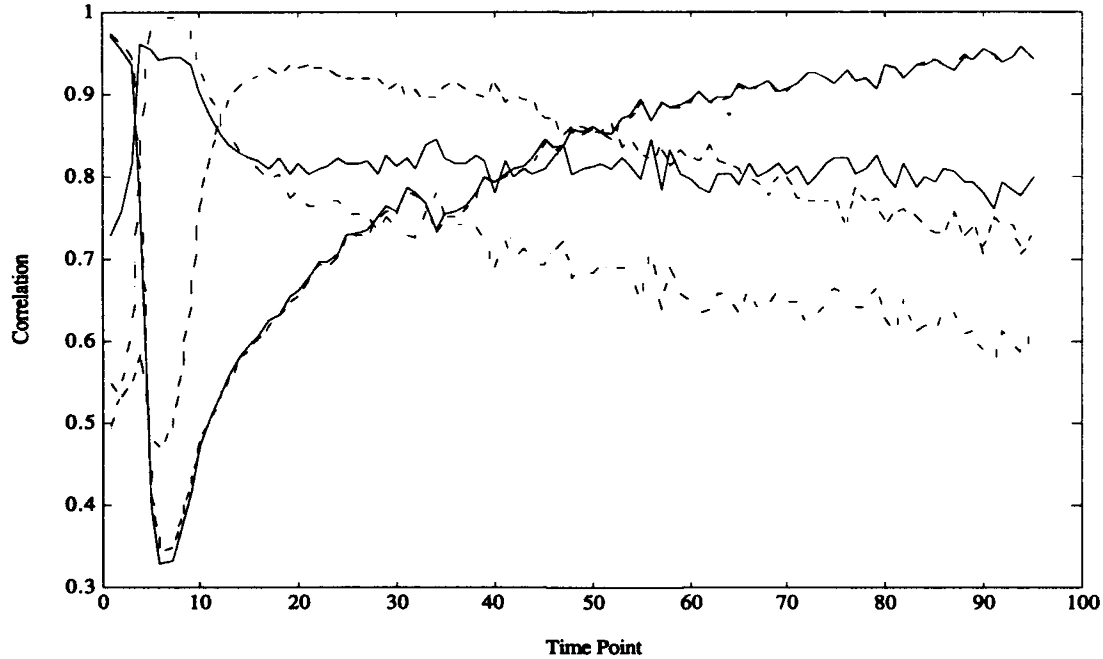


Figure 14. The correlation time-profiles of all the corners of the three-dimensional convex cone (defined by 6 extreme spectra for sample t15100p03).

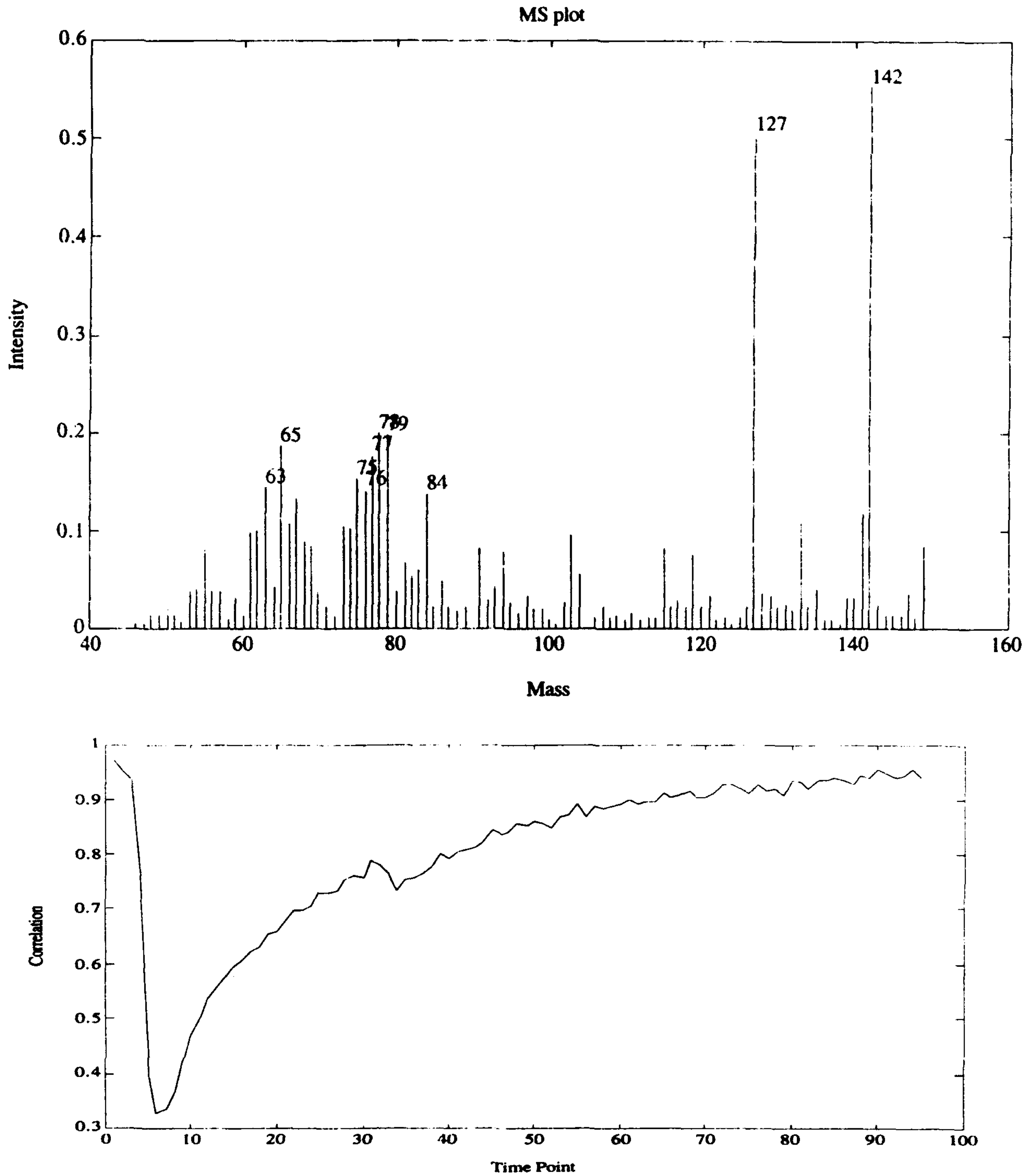
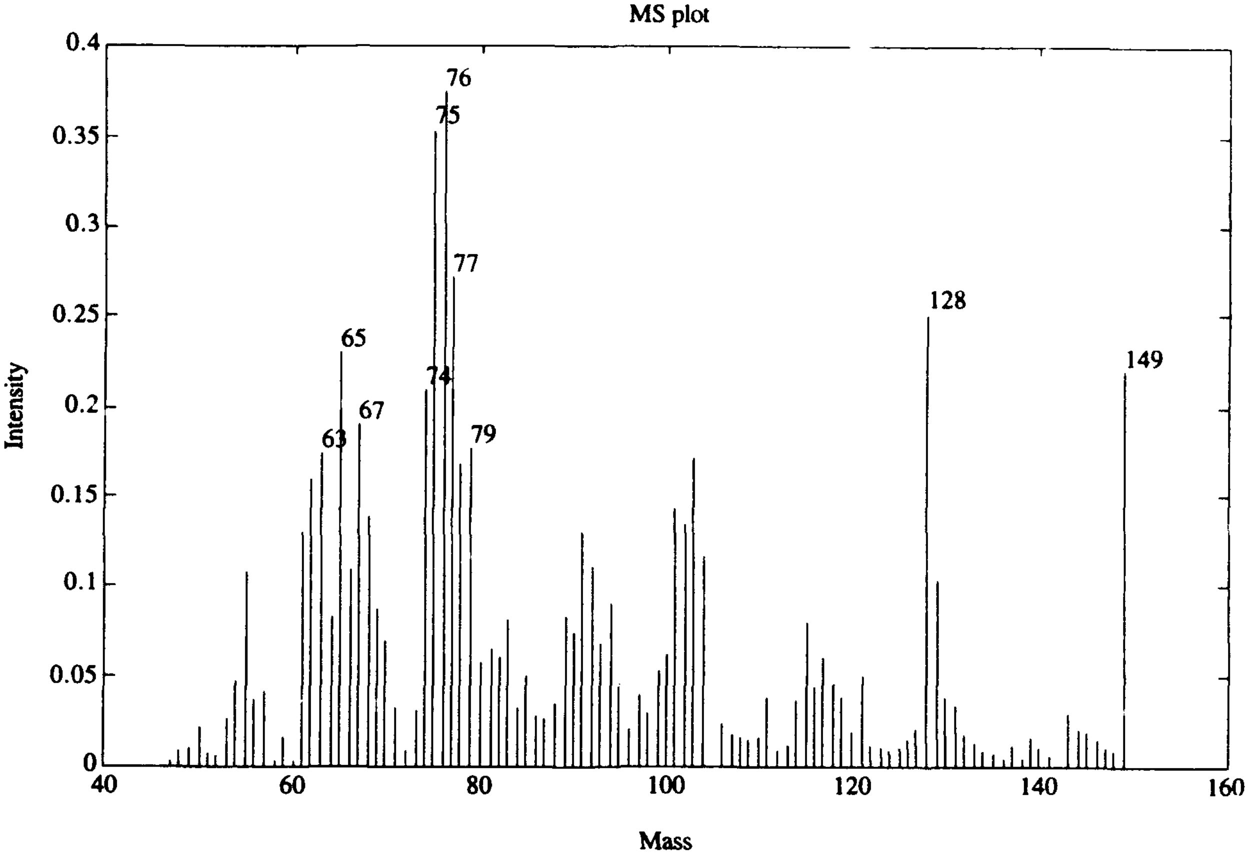


Figure 15. The spectrum and time-profile of the first extreme point of the three-dimensional convex cone of spectra (for sample t15100p03). This is the corner identified by a star at the left end of Figure 13.



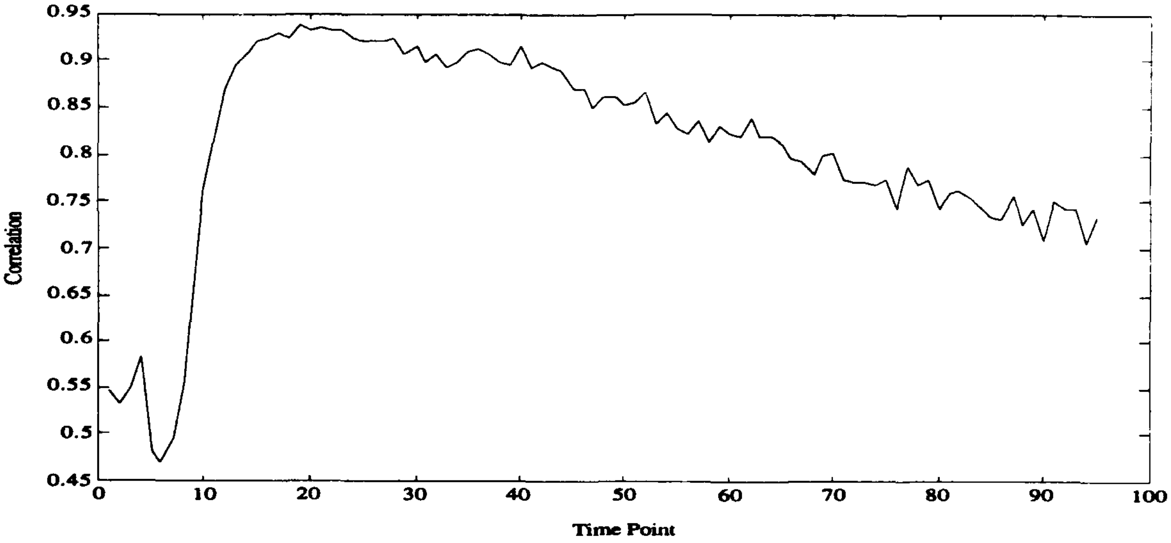


Figure 16. The spectrum and ti..ıe-profile of the second extreme point of the three-dimensional convex cone of spectra (for sample t15100p03). This is the corner identified by a circle at the bottom of Figure 13.

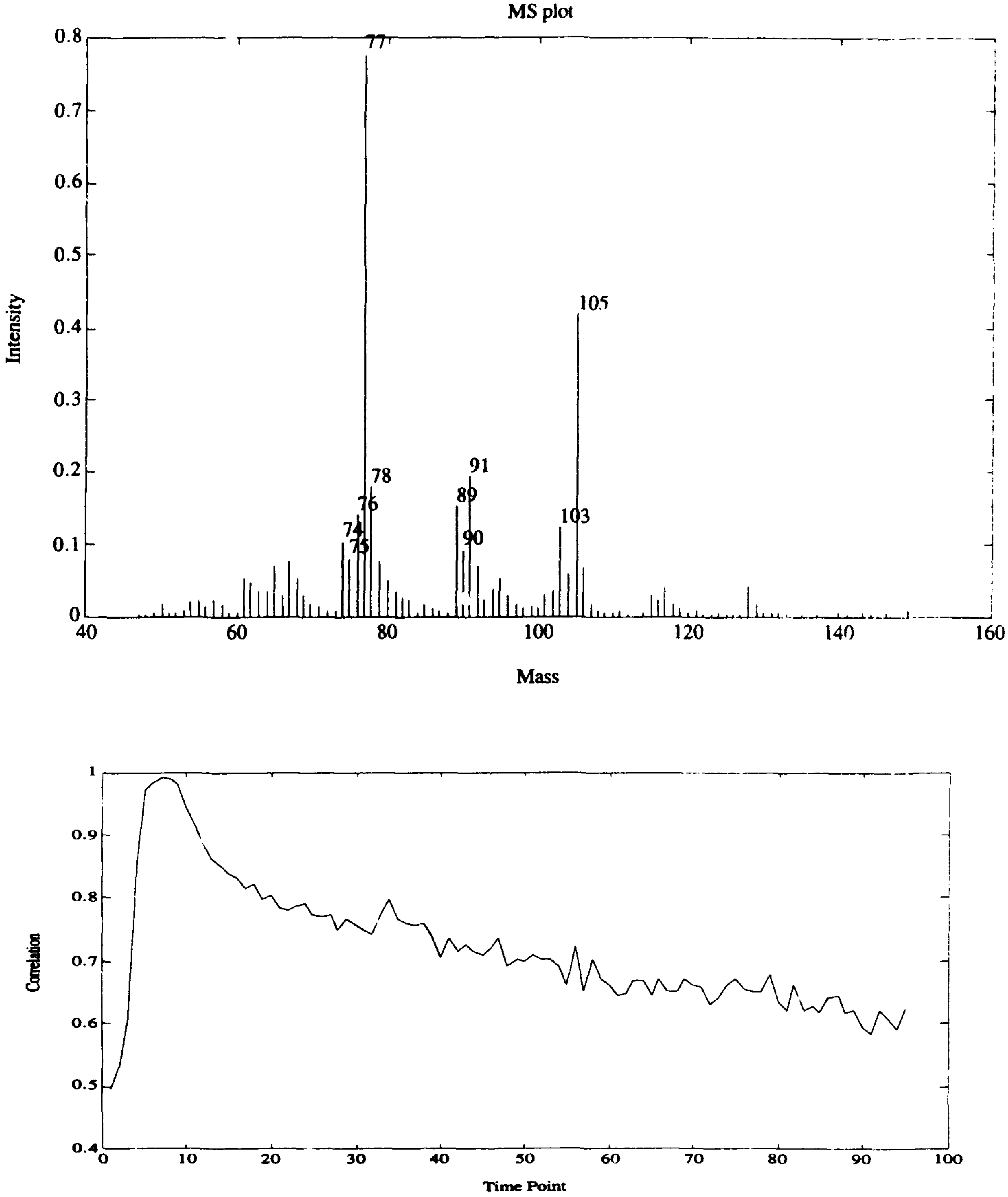


Figure 17. The spectrum and time-profile of the third extreme point of the three-dimensional convex cone of spectra (for sample t15100p03). This is the rightmost corner identified by a circle in Figure 13.

3.2. Convex Cone of Partial-Intensity Sample

In Figure 13, we construct the cone in the usual way, and we present the correlation time-profiles in Figure 14. In the irregular hexagon of Figure 13, we note that in addition to the formation of clusters of corners, we have a corner (with a value of 0.3 for Factor 2 and 0.5 for Factor 3) which forms a very obtuse angle. Clearly, elimination of that corner would cause only a minor modification of

the area and the shape of the polygon. Approaching the question from another viewpoint, the corner in question is very close to a positive linear combination of its two neighboring corners, and we can approximate it as such. In this (rather common) situation then we can simply ignore this corner and regard the polygon as virtual triangle.

The time profiles (Figure 14) very clearly have the three types of qualitative behavior we observed for sample t15100p08, with the exception of one profile (solid line with a maximum at time point 4) which behaves rather strangely. This is actually the obtuse-angle corner of Figure 13, discussed above, and its unusual behavior is a direct consequence of the fact that it is not a significant extreme point of the cone.

We show the spectra of only the first three corners (Figures 15 to 17), which happen to represent the three types. A comparison of Figure 15 to Figure 10, Figure 16 to Figure 6, and Figure 17 to Figures 7 to 9 shows that the decomposition of this partial sample corresponds closely to that of the full-intensity t15100p08, confirming the good sensitivity of this methodology.

Table 2. The 18 largest singular values for sample t15100p03.

σ_1 to σ_6	σ_7 to σ_{12}	σ_{13} to σ_{18}
9.3361	0.2287	0.1702
2.2180	0.2148	0.1628
1.3432	0.2004	0.1589
0.4827	0.1914	0.1544
0.3630	0.1850	0.1473
0.2500	0.1783	0.1422

4. BACKGROUND SAMPLE

We now consider the second pyrolysis cycle of Figure 1; this is sample t15100p02, shown in Figure 18. The total ion intensity shows that this pyrolysis cycle should not contain any biological material. However, we can draw this conclusion only because we can conveniently compare it to subsequent pyrolysis cycle. If we want to draw the conclusion at the time the pyrolysis cycle takes place, we must base it on the spectrum itself or on the time-profiles, without reference to the absolute intensity. This is clearly a non-trivial task.

4.1. Principal Components of Background Sample

Extraction of the first three factors is carried out in the usual way, and their correlation time-profiles are shown in Figure 19. It is worth noting that these profiles have more noise than those of the biological sample (Figure 3).

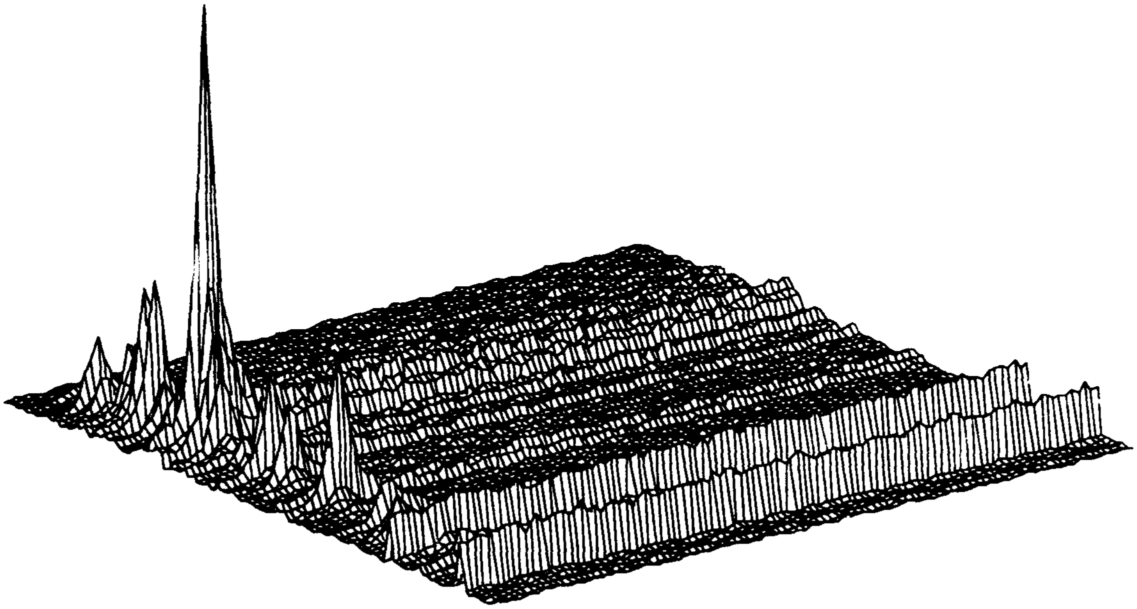


Figure 18. The spectrum for file t15100p02. The time axis is oriented from the front corner towards the right. The mass axis is in the front left.

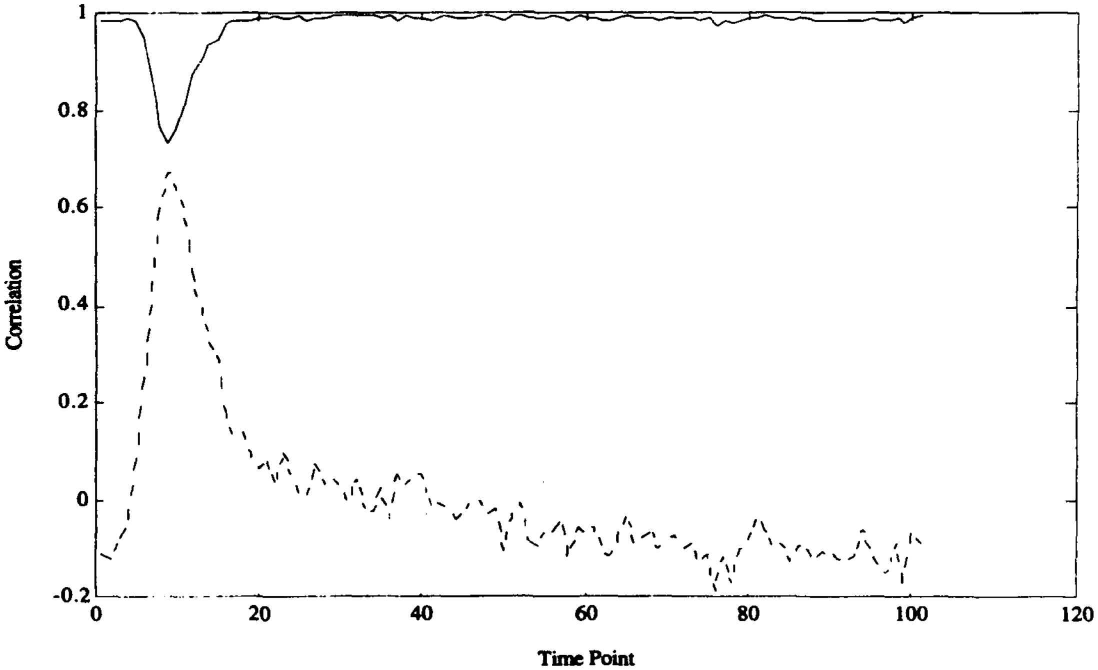


Figure 19. Time profiles for the first three principal components of t15100p02.

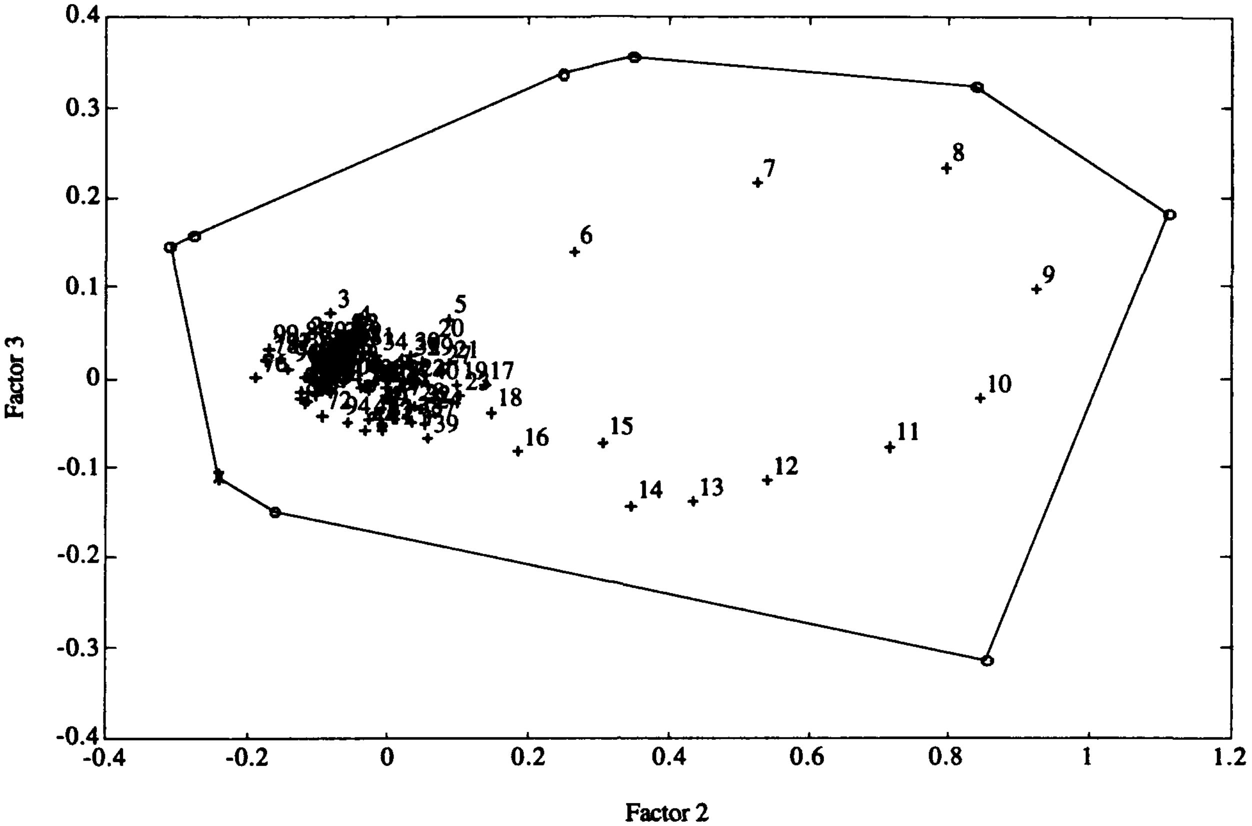


Figure 20. The three-dimensional cone of spectra for t15100p02 (shown in the form of the coefficients of Factor 2 and Factor 3, while the coefficient of Factor 1 is preset to 1).

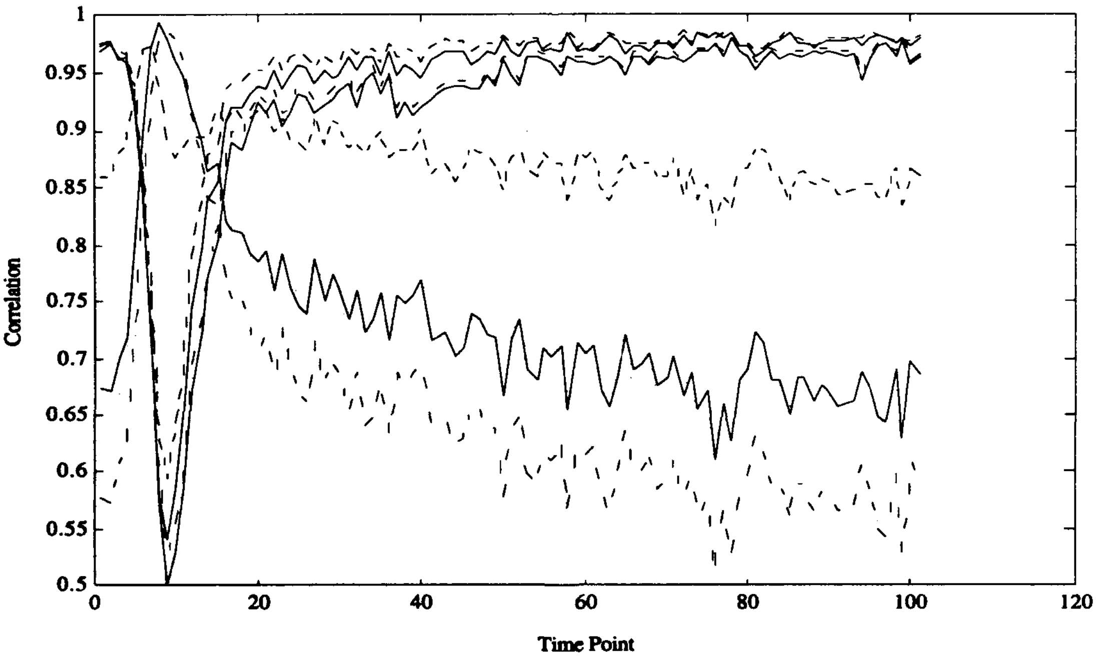


Figure 21. The correlation time-profiles of all the corners of the three-dimensional convex cone (defined by 9 extreme spectra for sample t15100p02).

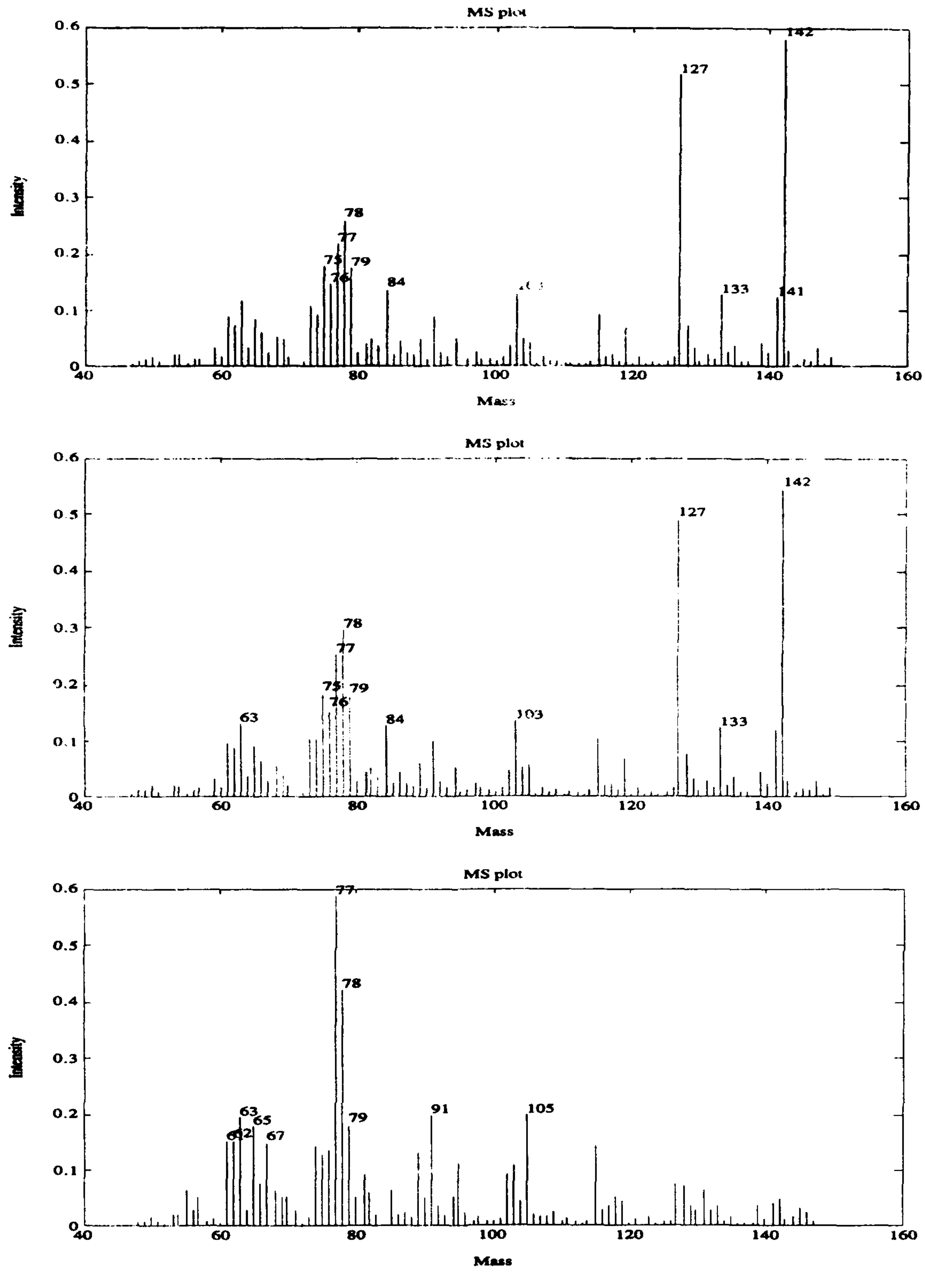


Figure 22. Spectra of corners 1-3 (counting counterclockwise from the star corner in Figure 20) for the three-dimensional convex cone of t15100p01.

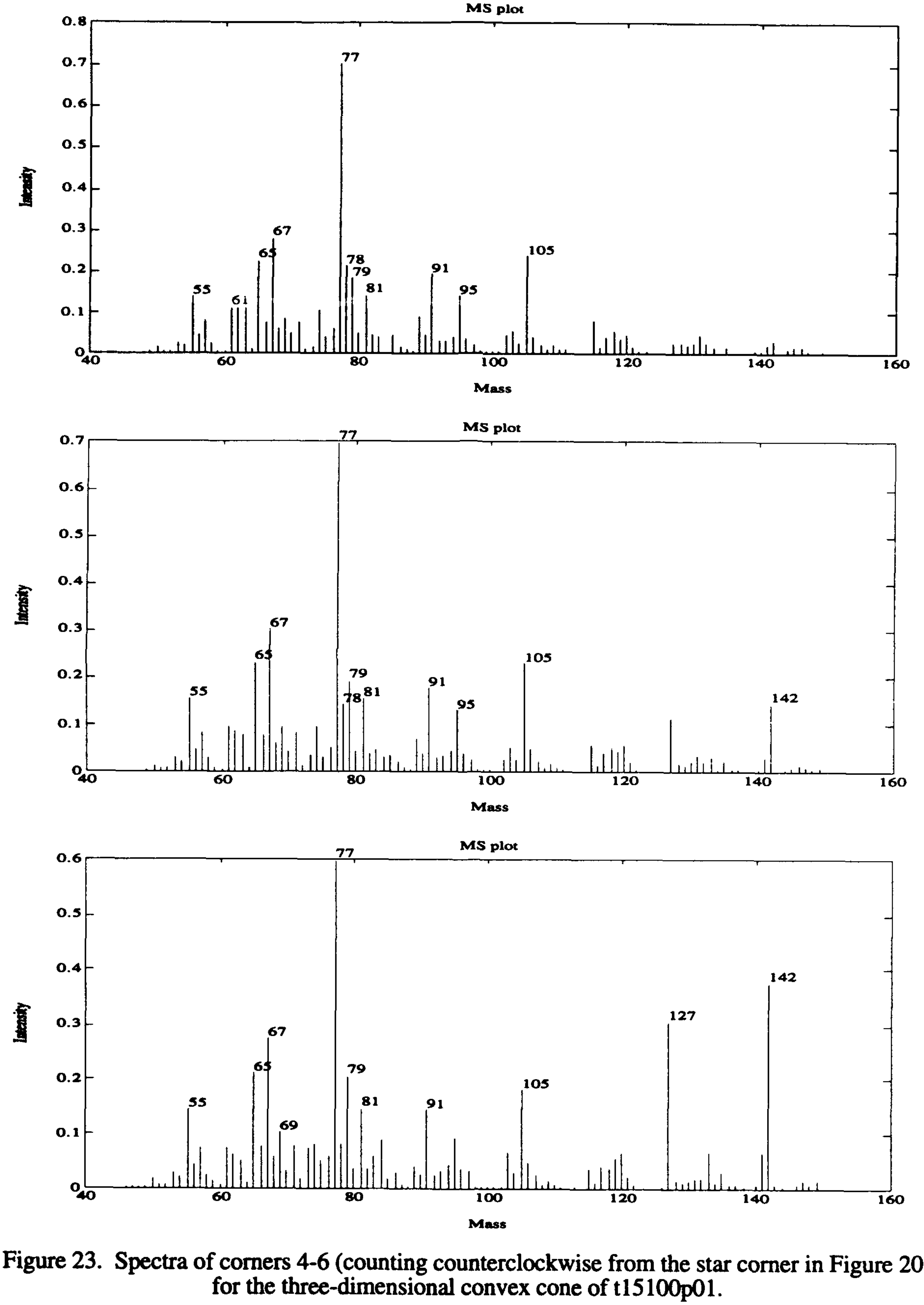


Figure 23. Spectra of corners 4-6 (counting counterclockwise from the star corner in Figure 20) for the three-dimensional convex cone of t15100p01.

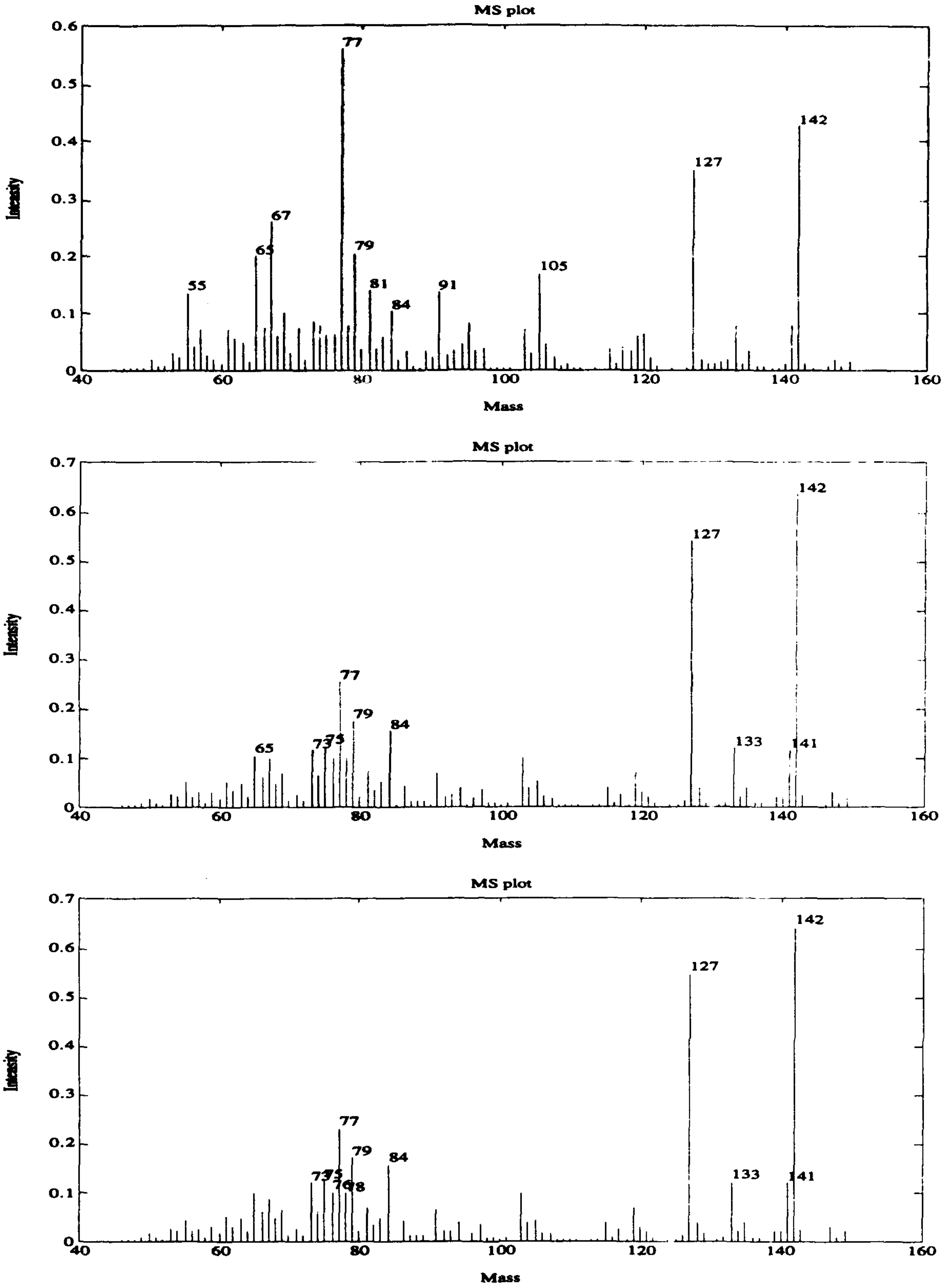


Figure 24. Spectra of corners 7-9 (counting counterclockwise from the star corner in Figure 20) for the three-dimensional convex cone of t15100p01.

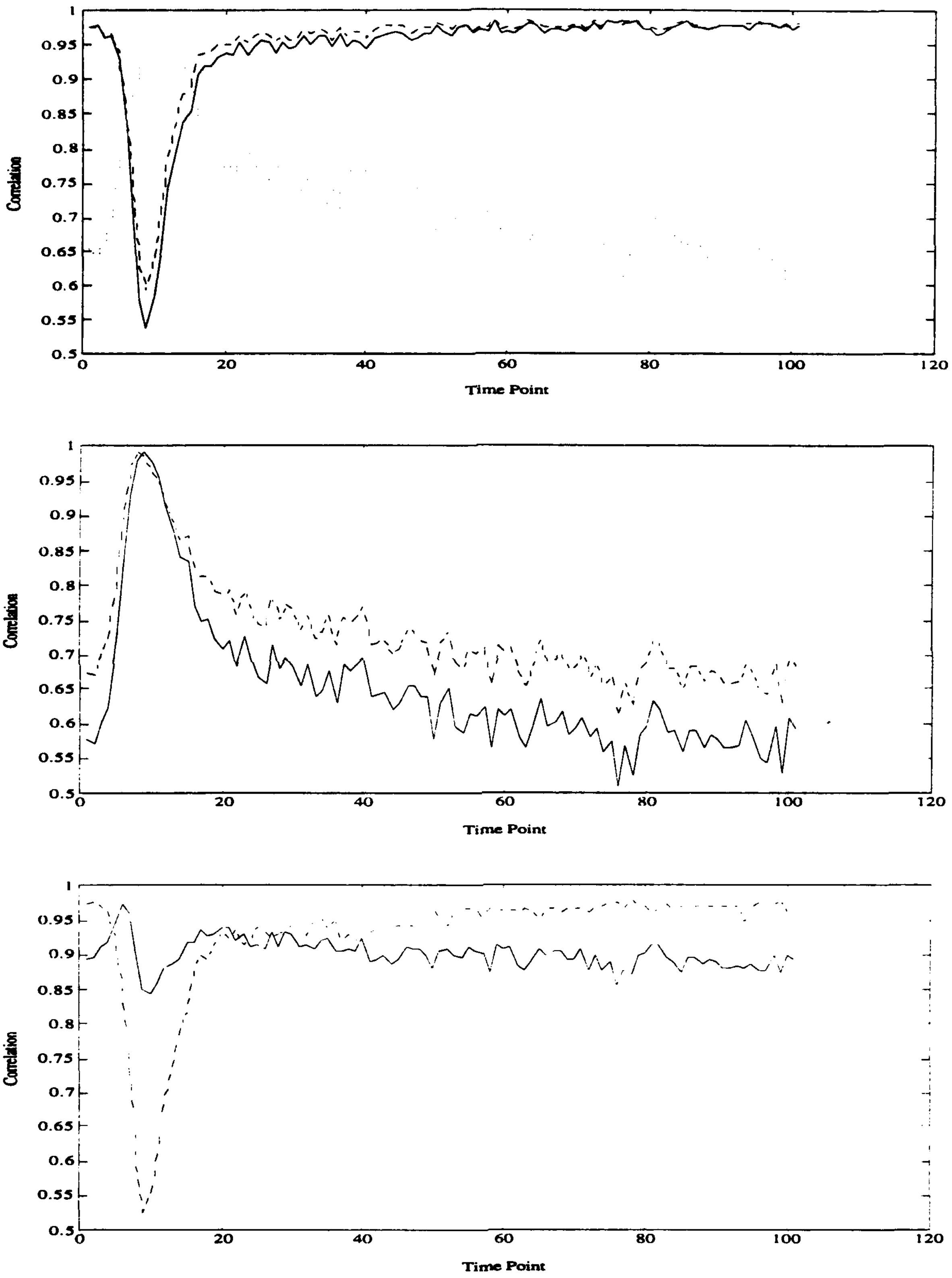


Figure 25. Time-profiles of corners (enumerated counterclockwise from the star corner in Figure 20) for the three-dimensional convex cone of t15100p01.
Top: Corners 1-3. Middle: Corners 4-6. Bottom: Corners 7-9.

4.2. Convex Cone of Background Sample

Construction of the three-dimensional convex cone (Figures 20 and 21) produces 9 corners - many more corners than the biological sample (Figure 4), which had only 5. It is also apparent that the shape in Figure 20 cannot be easily reduced to a triangle.

In Figure 21, we note that no time-profile exhibits clear dominance in the region of time-points 15-20. In fact, those profiles that have clear extrema are either of the background-characteristic form (maximum in the vicinity of time point 8) or have the corresponding echo behavior (minimum in the vicinity of time point 8).

Corners 1, 2, and 3 (the lowest three in Figure 20) would be the most promising to contain biological-material information if we consider the location of the time points 15-20; but note that 20 actually is practically in the center of a region that includes corners 6, 7, 8, and 9 (the 4 corners in the top left quadrant of Figure 20), further demonstrating the absence of biological sample information.

Therefore, we can discern a distinctly different behavior (and one that could be used to distinguish background from the presence of biological material) in the behavior of the cone in Figures 20 and 21, without even examining the actual mass spectra of the corners!

This conclusion can be verified through a closer examination of the spectra of the corners, shown in Figures 22 to 24, and the corresponding time-profiles (Figure 25). We note that in Figure 25 there seem to be some promising time-profiles, in terms of shape. But their correlation is very noisy and it never rises above the echo profiles in the region of interest.

5. CLASSIFICATION

5.1. Samples Used for Classification

In its final part, this study considered the automatic clustering of samples into classes, through simple algorithms.

Three categories of samples were involved:

(A) Egg Albumin

(B) *Bacillus Subtilis* Var. *Niger (Bacillus Gobulii)*

(C) *MS-2 Coliphage*

The data sets used for the study were those prefixed by t151** and t152**. Their total ion intensity profiles are shown in Figure 1 (for t15100) and Figures 26-29.

As we discussed earlier, pyrolysis cycles without biological material can be identified based solely on the shape of the cone and the absence of suitable time-profiles of corners. Thus, in this section we will focus on the classification of the biological materials, rather than the detection of biological material vs. background.

We will label samples with the letters A, B, and C, according to their category, but we will use lower-case letters for samples which, based on the total intensity profiles (Figures 1 and 26-29), do not have a full amount of biological material sample. Based on this convention, the pyrolysis cycles of t15100 (Figure 1) are labeled as 00cCCCCC. The captions of Figures 26-29 show the classification labels of the other samples.

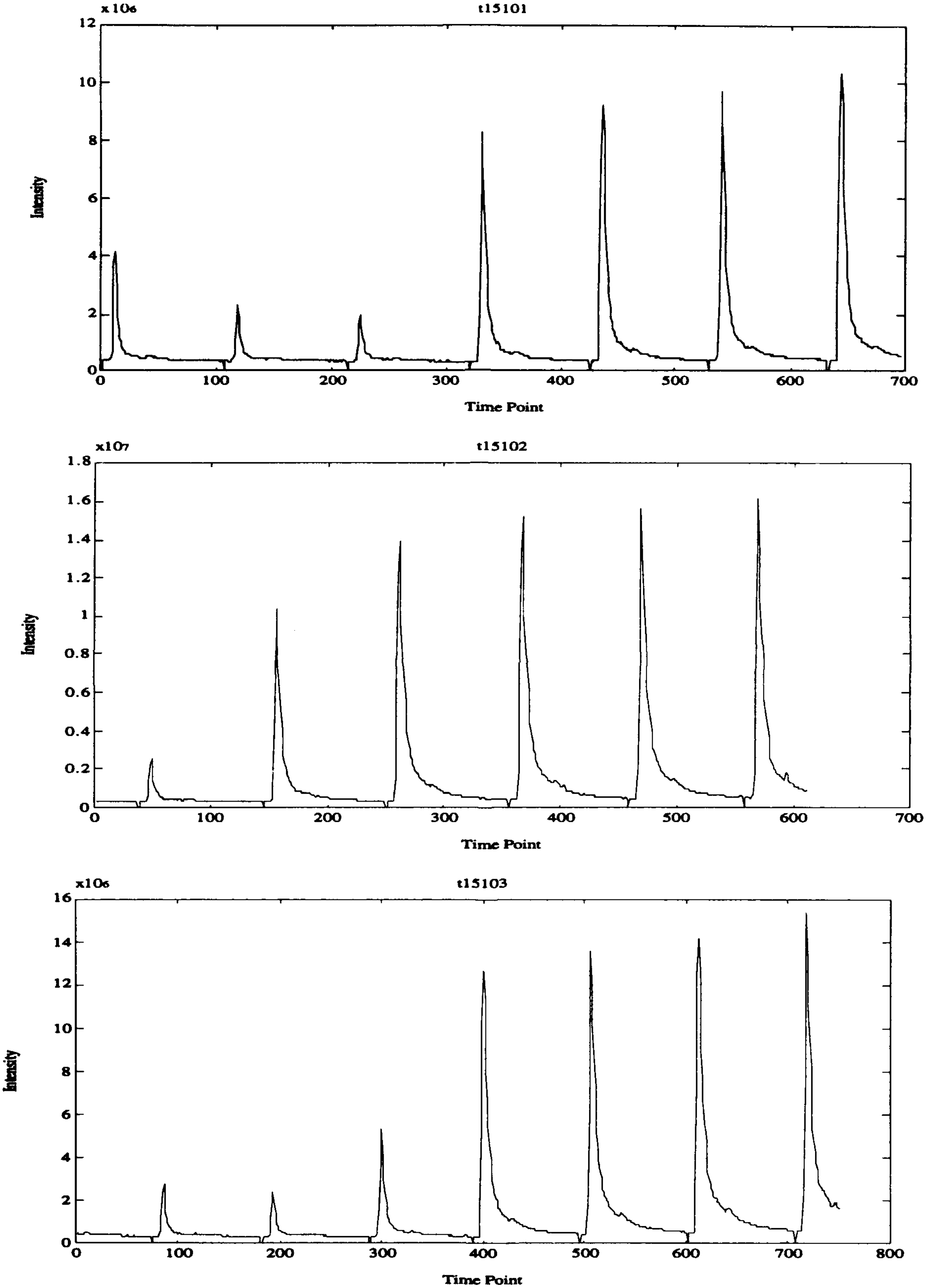


Figure 26. The total ion intensity profiles for samples t15101 (classified as 000aAA), t15102 (classified as 0bBBB), and t15103 (classified as 00cCCC).

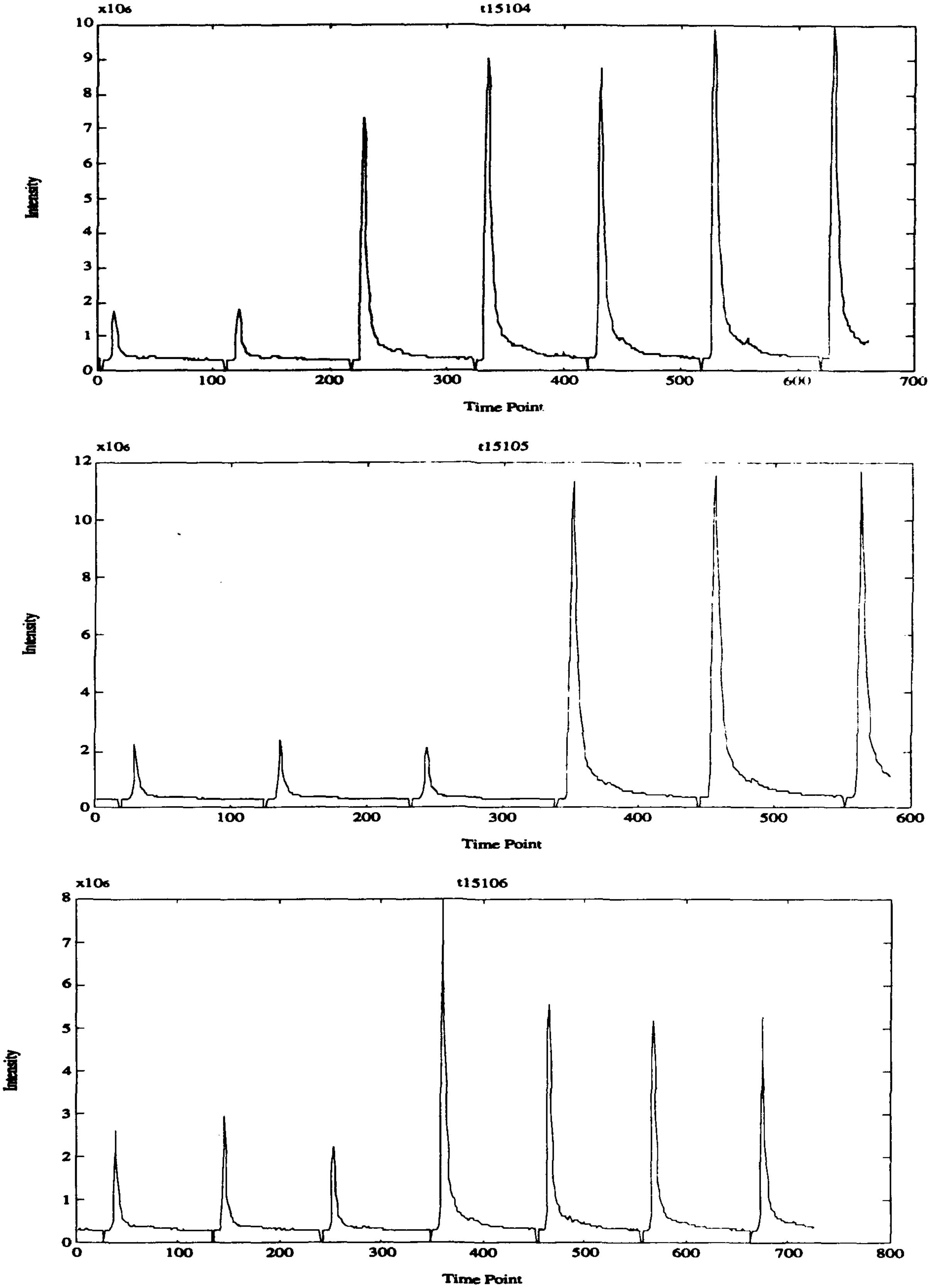


Figure 27. The total ion intensity profiles for samples t15104 (classified as 00aAAA), t15105 (classified as 000BB), and t15106 (classified as 000BBB).

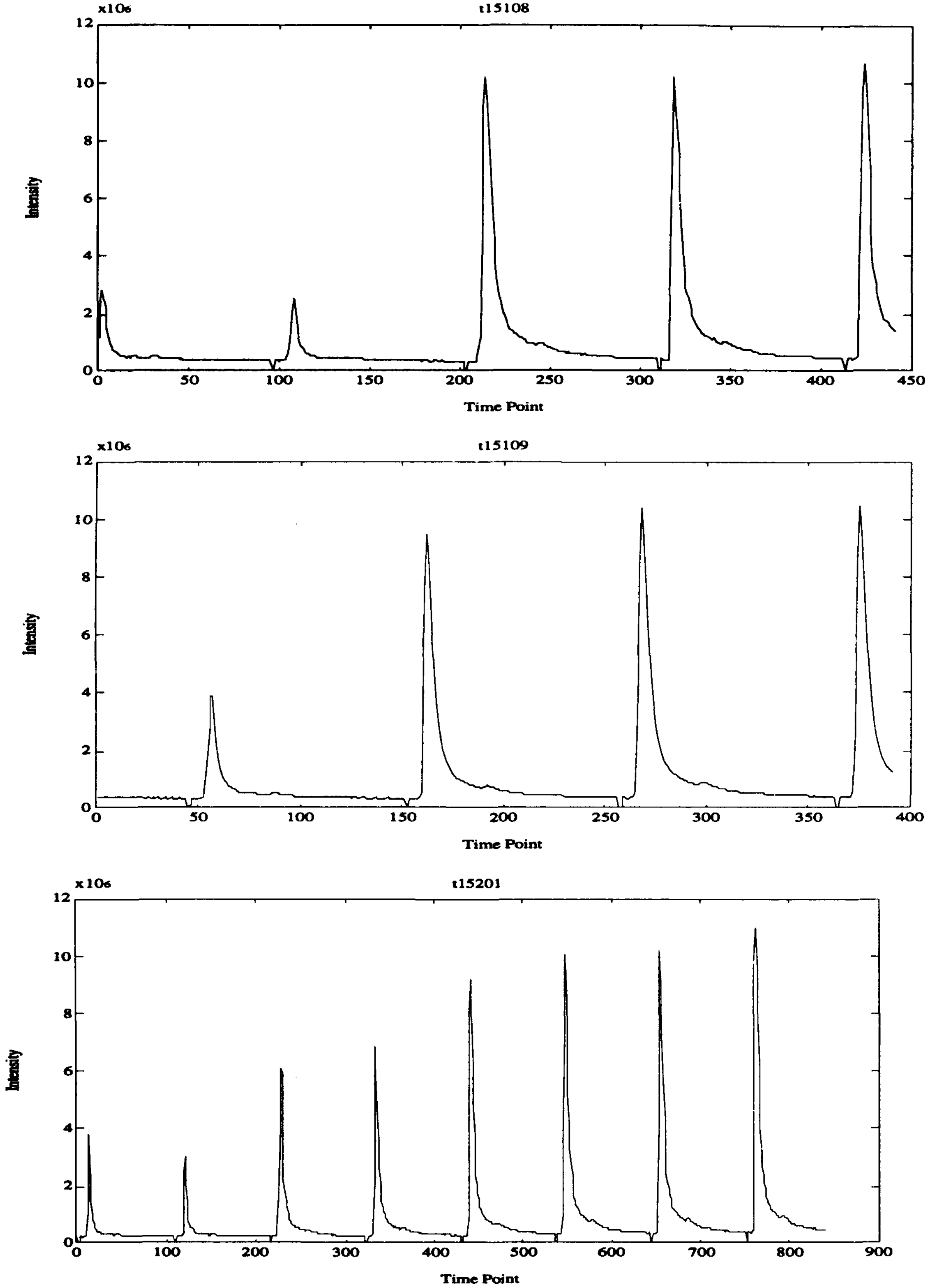


Figure 28. The total ion intensity profiles for samples t15108 (classified as 00CC), t15109 (classified as 0aA), and t15201 (classified as 00ccCCC).

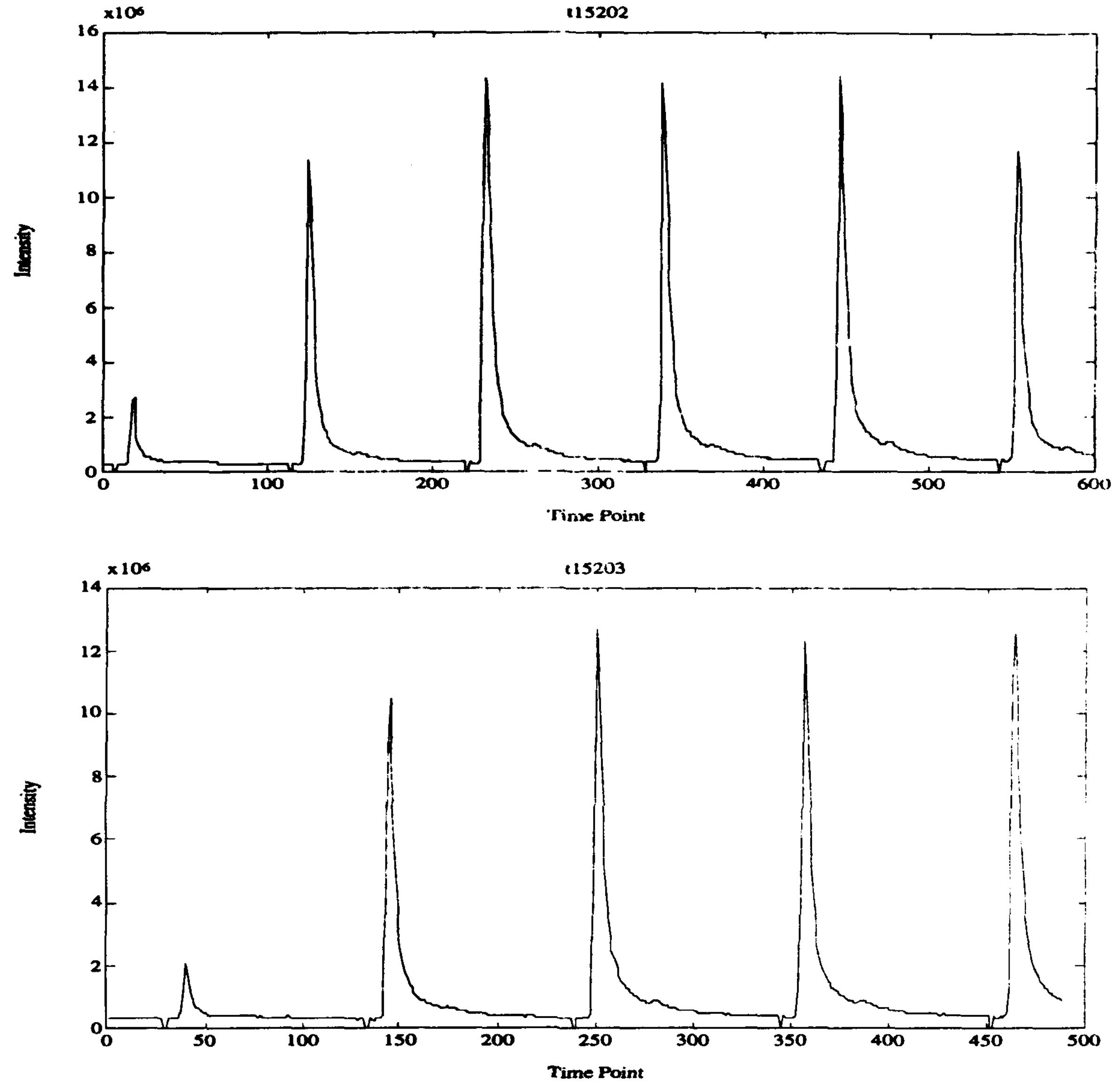


Figure 29. The total ion intensity profiles for samples t15202 (classified as OaAAA) and t15203 (classified as ObBB).

5.2. Selection of Corners

The corner that is most representative of the biological material contained in the sample has, as was discussed earlier, a maximum correlation in its time-profile in the vicinity of time points 15 to 20 with a gradual decline thereafter; furthermore, the gap between these profiles and profiles of the other two types is quite large at time point 20. Naturally, two or three corners might exist with this kind of behavior, but their spectra would then be similar, so that no need would exist to exercise care in the selection of one corner from this group.

Since the differences are so drastic, we select a corner for each sample simply be examining the correlation value at time point 20. Thus, each sample is represented in the classification by the spectrum of the corner that has the highest correlation at time point 20.

5.3. Classification Technique

Given the exploratory character of this study, we will attempt here a clustering of the samples without using our prior knowledge of the correct labels. We will carry out a very simple form of clustering using the spectra (but not the time-profiles) of the corners.

Spectra are represented as normalized 104-dimensional vectors of intensities. Our measure of proximity of two spectra will simply be the inner product of the vectors (since they are normalized, this quantity corresponds to the cosine of the angle formed by the two vectors). We can thus construct a matrix of pairwise distances, which, if A is a matrix that contains as its columns the sample spectra, is simply the covariance matrix A^TA.

We view the samples as nodes in a complete graph (i.e., a graph where all pairs of nodes are connected by an edge), pairwise distances are assigned as weights on these edges. In this context, we view the clustering as a construction of a spanning tree, i.e., a tree (a connected graph without cycles) that includes all the nodes of the graph. We seek the minimum spanning tree, i.e., the spanning tree whose sum of edge weights is as small possible.

This tree can be constructed using Kruskal's algorithm (see, for example, Sedgewick, 1988, p. 458-461; Roberts, 1984, p.526-529) as follows: We select the edge with the smallest weight and include it in the tree. We then examine other edges consecutively, ordered by decreasing weight, and we include an edge in the tree if it will not form a circuit (a closed path) with the edges already existing in the tree.

5.4. Resulting Spanning Trees

5.4.1. Complete Set of Samples

We carry out the construction of the spanning tree for the data set described above, and, in order to visualize the tree, we also carry out SVD of the matrix of spectra. The results are shown in Figures 30 and 31, which give all three possible two dimensional projections of the tree using (two of) the first three Factors.

We should emphasize that the tree itself is constructed in the full 104-dimensional space of the spectra, and is afterwards projected onto the two-dimensional subspaces of Figures 30 and 31, purely for visualization purposes. Therefore proximity in the figures does not necessarily imply proximity in the full space or in the construction of the tree.

Figure 31 and, to a less clear form, Figure 30 show excellent separation of the classes, which appear to form natural subtrees, even though no information on the label of each sample was used in the construction of the tree. We note that partial-intensity samples (lowercase letters in the figures) are usually at the outer branches of the subtree of each class, but they are nevertheless part of the correct subtree.

One exception to the last observation is sample t15201p03 (partial-intensity sample labeled c) which appears to be an outlier. It is connected to class B, but it is in reality a class of its own (note especially its drastically different Factor 1).

Note, finally, that even though the tree was constructed in a high-dimensional space, there is considerable geometric separation of the samples in the space of the first few factors.

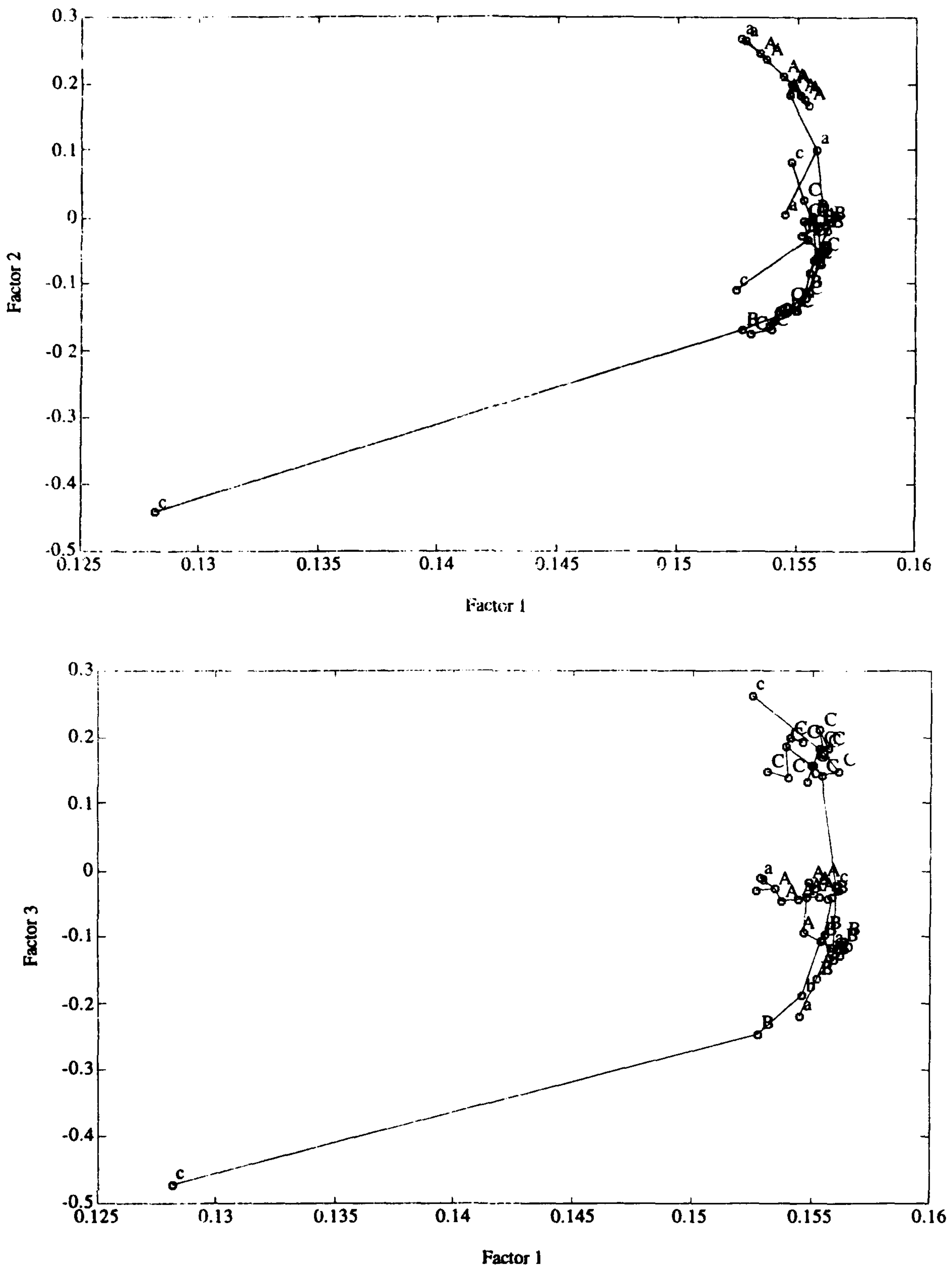


Figure 30. Spanning tree of samples, using Factor 1.

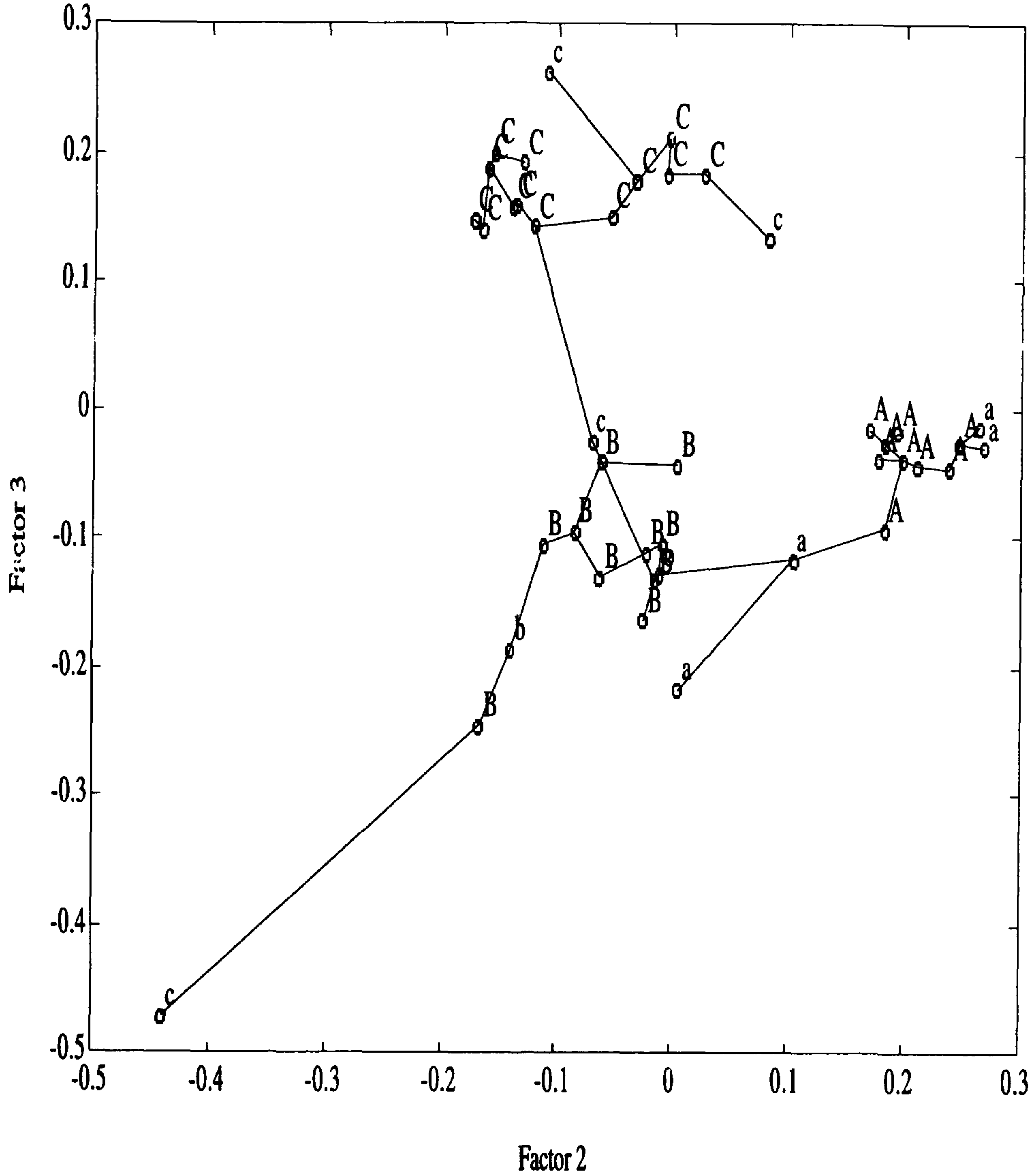


Figure 31. Spanning tree of samples, in Factor 2 / Factor 3 space.

5.4.2. Removal of Outlier

Given the strong bias introduced by the outlier, we will recompute the tree and the SVD after exclusion of the outlier. We then return to identify the reason the causes t15201p03 to be misplaced.

Figures 32 and 33 show the new tree, in the same projections as before, for the new data set. The distinct subtrees for the three categories, as well as the clear separation in the two dimensional projection are now much clearer. As before, partial-intensity samples (lowercase letters) are correctly clustered but usually lie in the outer branches of the subtree of any given class.

5.4.3. Cause of Outlier

Returning to the origin of the outlier, we examine the convex cone generated by t15201p03 (Figures 34 and 35), and in particular corners 2 and 3 (where the numbering begins with corner marked by a star and proceeds counterclockwise); these are shown in Figures 36 and 37. We note that corner 2 has the desired profile.

Corner 3 has a much different spectrum (Figure 37 vs. Figure 36) but it is close to the line connecting corners 2 and 4; through a coincidence in the time profiles, corner 3 actually has a higher correlation than corner 2 at time point 20. We therefore conclude that the simple-minded manner in which we picked the corner relevant for biological material characterization has not worked well in this example. This occurrence shows the need for systematic conversion of the polygon into a triangle.

6. CONCLUSIONS

In summary, this work addressed the classification of high-dimensional time-dependent pyrolysis mass spectra of biological samples. The data were projected onto a *low-dimensional subspace* using singular value decomposition. A convex cone was then formed on this subspace showing, as its corners, physically meaningful components of the sample. The convex cone contains only the physically meaningful spectra of the subspace. This technique enabled separation of the signal most characteristic of the biological material from the background signals and the interference of the device. The biological-material signal could be identified independently of the absolute amount of sample.

In our analysis of a pyrolysis cycle before the presence of the biological sample, we showed that the detection of the presence of any biological material could be accomplished based on the convex cone alone, without other reference to the mass spectra. This is possible because in the absence of a material sample the cone has many more corners, and the time-profiles of the corners do not exhibit the qualitative behavior (specifically, a maximum in a particular time region) that is expected from biological samples.

After the extraction of the appropriate corner of the convex cone, we carried out automatic clustering of the samples, by using a minimum spanning tree algorithm with pairwise distances. We showed that full-intensity and partial-intensity samples cluster correctly.

The technique of convex cones is a very promising approach for analyzing and classifying high-dimensional and/or time-evolving spectra.

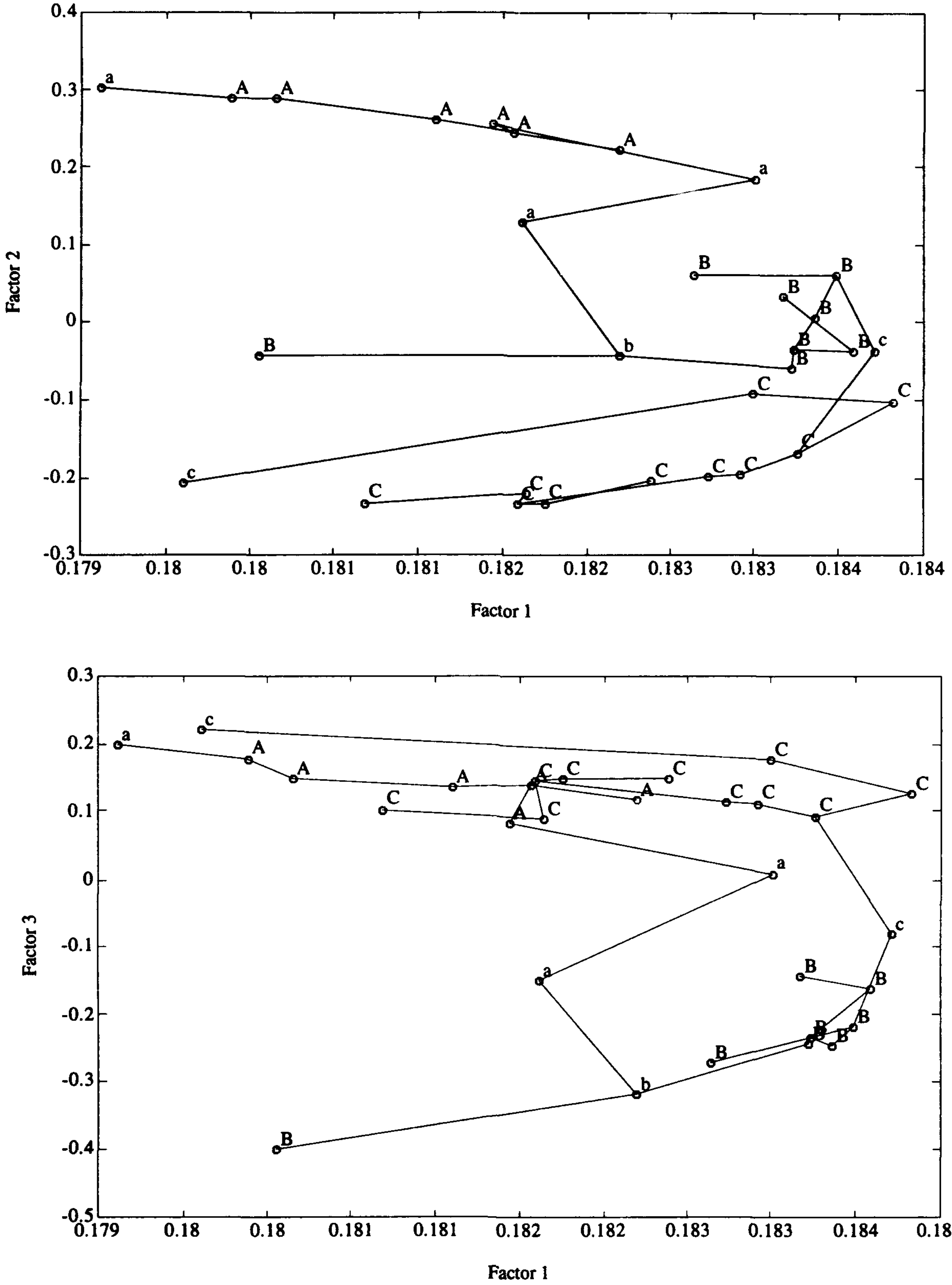


Figure 32. Spanning tree of samples after removal of outlier t15201p03, using Factor 1.

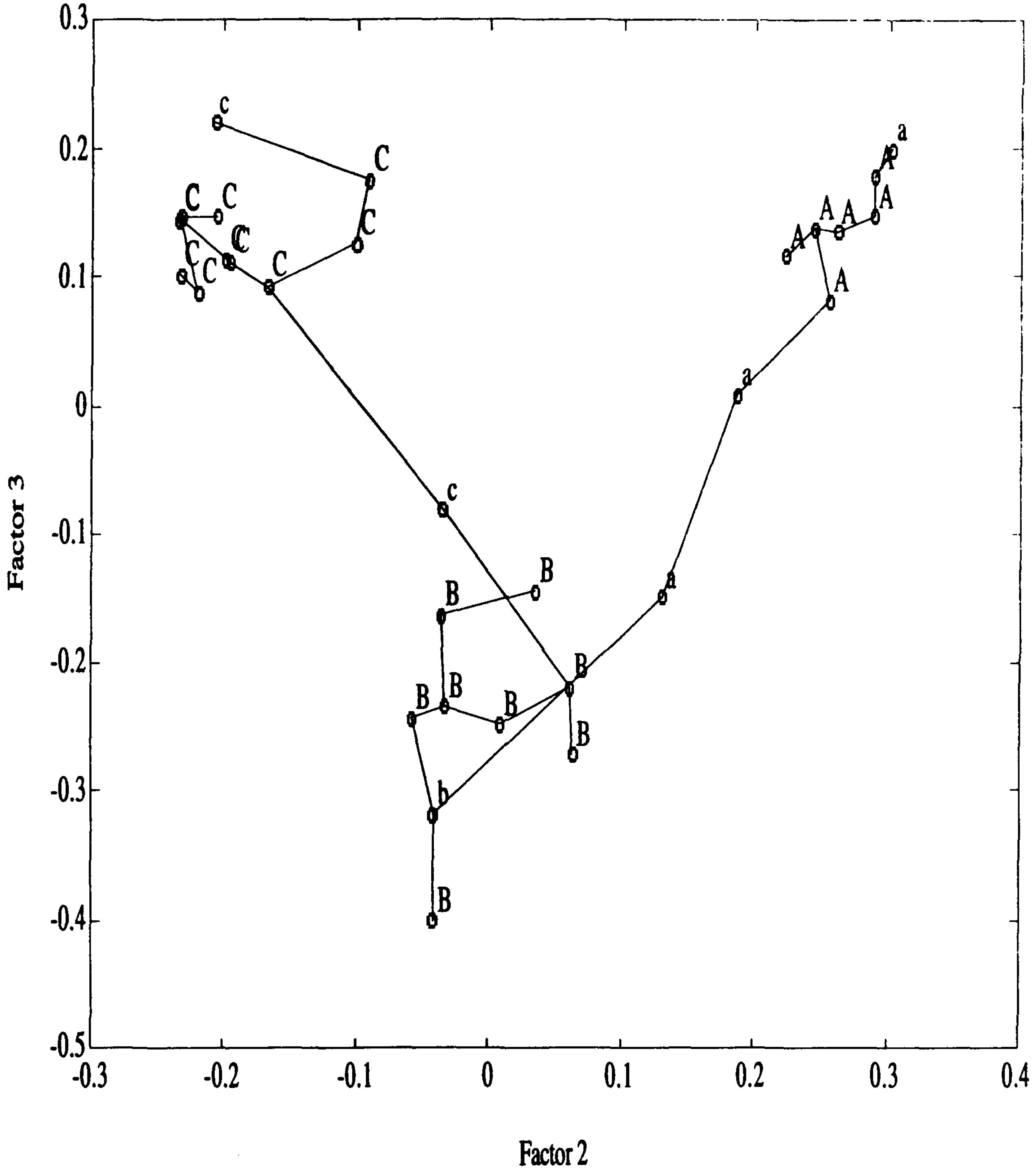


Figure 33. Spanning tree of samples after removal of outlier t15201p03, in Factor 2 / Factor 3 space.

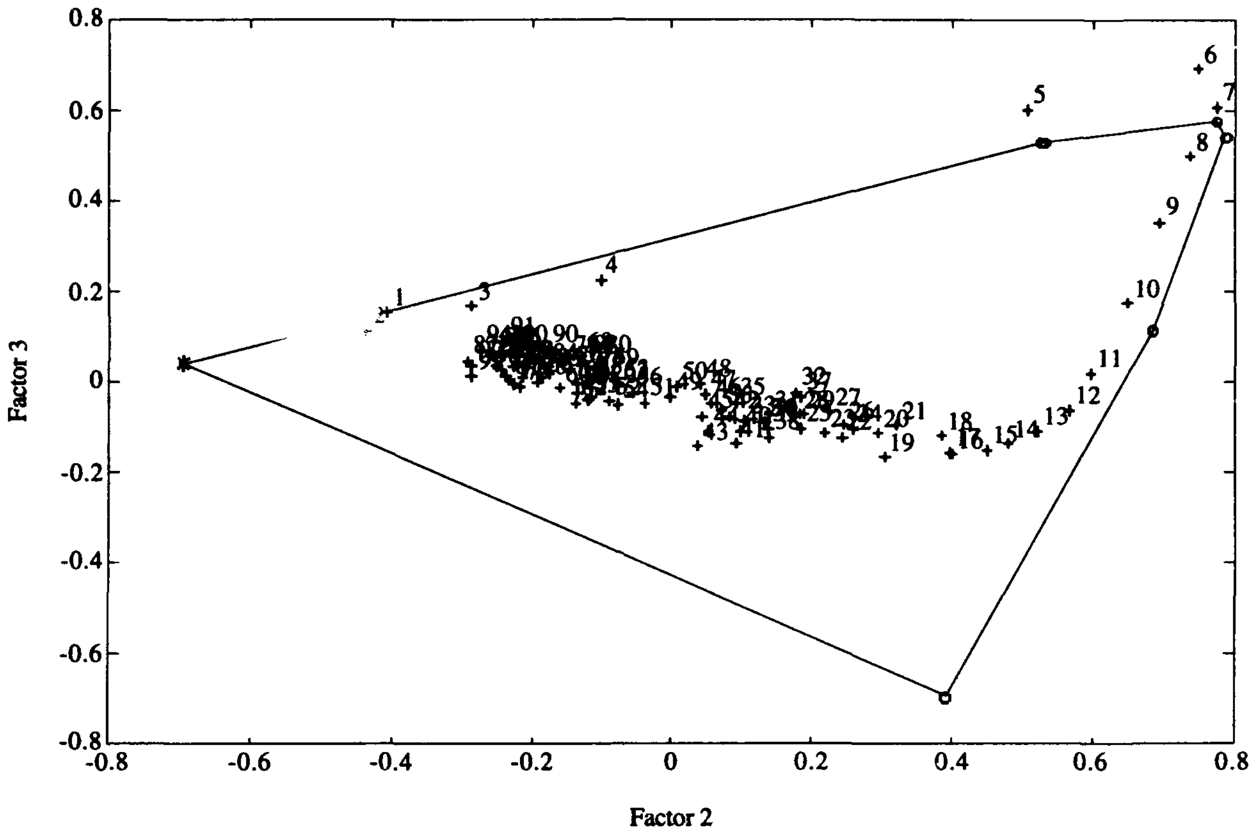


Figure 34. The three-dimensional cone of spectra for t15100p02 (shown in the form of the coefficients of Factor 2 and Factor 3, while the coefficient of Factor 1 is preset to 1).

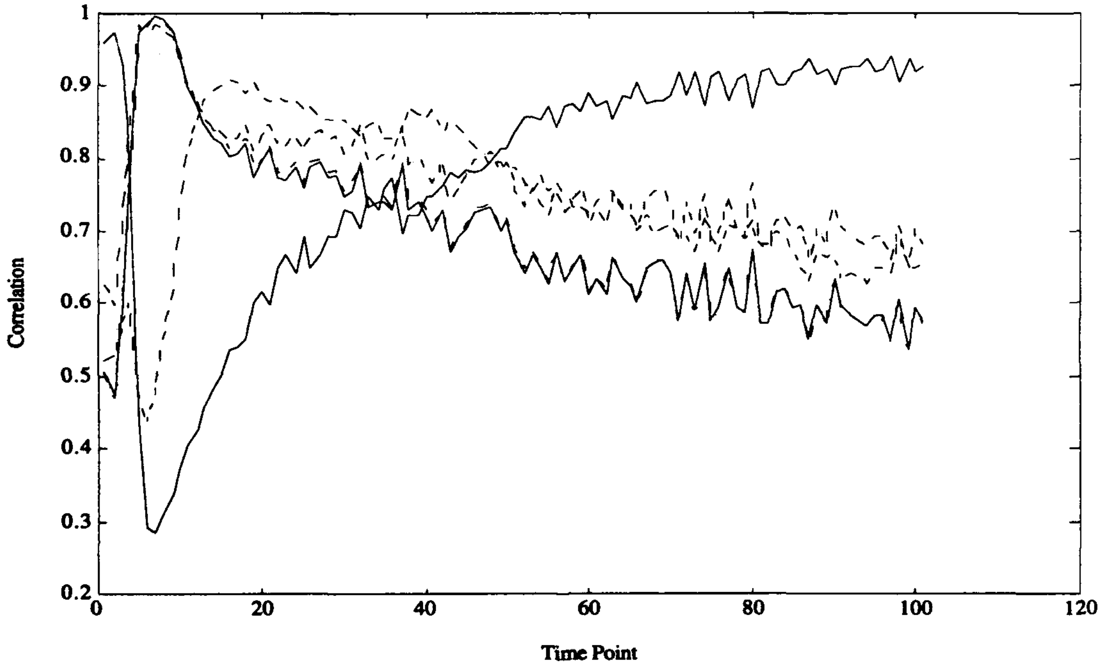


Figure 35. The correlation time-profiles of all the corners of the three-dimensional convex cone (defined by 6 extreme spectra for sample t15100p02).

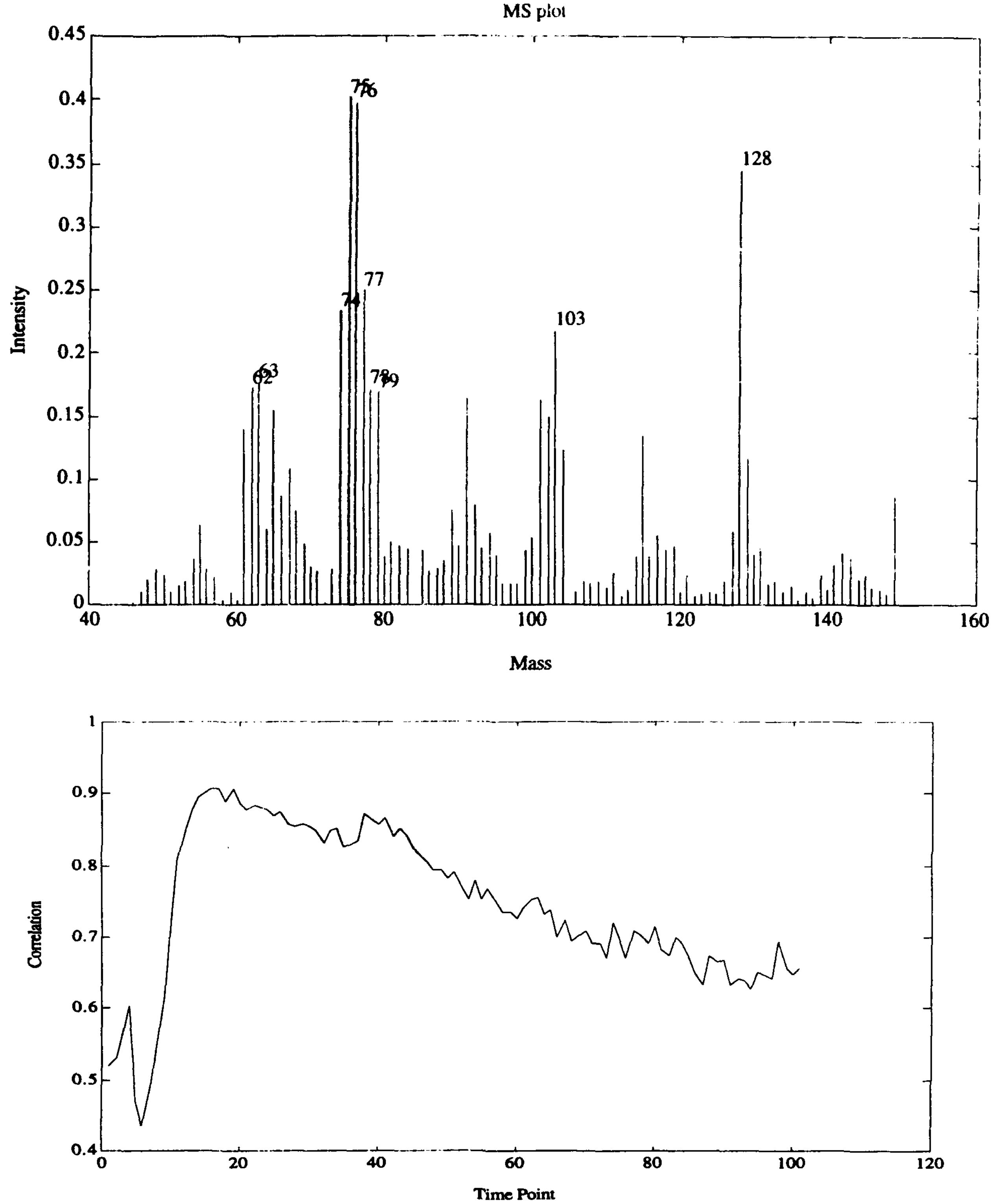


Figure 36. The spectrum and time-profile of the second extreme point of the three-dimensional convex cone of spectra (for sample t15201p03). This is the corner identified by a circle at the bottom of Figure 34.

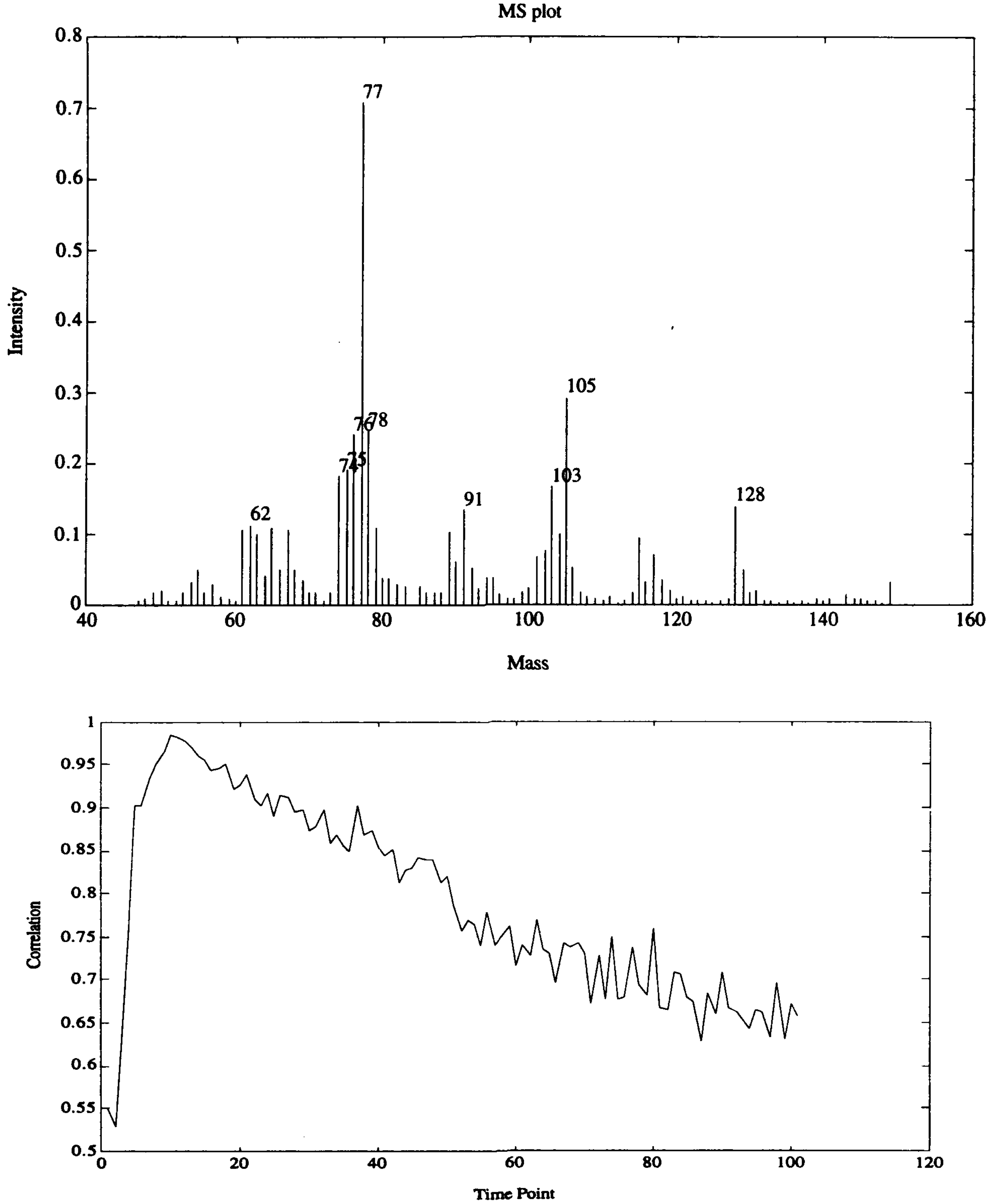


Figure 37. The spectrum and time-profile of the third extreme point of the three-dimensional convex cone of spectra (for sample t15201p03). This is the corner with coordinates of approximately (0.7, 0.1) in Figure 34.

Blank

REFERENCES

Harper, A.M., and Mavrovouniotis, M.L. "Convex-Cone Analysis of the Time Profiles of Pyrolysis Mass Spectra of Biological Agents" Technical Report, U.S. Army Chemical and Biological Defense Agency, 1993.

Ingraham, J.L., Maaløe, O., and Neidhardt, F.C. *Growth of the Bacterial Cell.* Sinauer Associates, Sunderland, Mass., 1983.

Jolliffe, I.T. *Principal Component Analysis.* Springer-Verlag, New York, 1986.

Sedgewick, R. *Algorithms.* Addison-Wesley, Reading, Mass., 1988.

Lehninger, A.E. *Principles of Biochemistry,* Worth Publishers, New York, 1982.

Roberts, F.S. *Applied Combinatorics.* Prentice-Hall, Englewood Cliffs, 1984.

Sickenberger, D., Ifarraguerri, A., Williams, R., Harper, A., Robbins, R., and Sarver, E. "Chemical Biological Mass Spectrometer" U.S. Army Chemical Research, Development, and Engineering Center, Aberdeen Proving Ground, 1992.

Strang, G. *Linear Algebra and its Applications, 3rd edition,* Academic Press, 1988.

Voorhees, K.J., DeLuca, S.J., Noguerola, A. "Identification of chemical biomarker compounds in bacteria and other biomaterials by pyrolysis-tandem mass spectrometry", *Journal of Analytical and Applied Pyrolysis*, **24**, p. 1-21, 1992.

Blank

APPENDIX

MATLAB SCRIPTS AND FUNCTIONS

The scripts and functions given here work in conjunction with those listed by Harper and Mavrovouniotis (1993). Script files are preceded by a line with the word "SCRIPT" and the name of the script, but this line was added for the purposes of identifying the script, and it is not present in the MATLAB file itself.

```
function [t, history]=clustering(a, c, plottree, plotgraph, ids)

% Forms a spanning tree, picking out maximal correlations.
% Input a is a matrix of pairwise distances of n points.
% Optional input c is a nx2 matrix of coordinates of the points,
%  to be used in plotting the tree.
% Plotcodes determine what kind of plotting will be done.
%  plottree=0 no plotting of tree (default)
%  plottree=1 plot edges of tree without pausing
%  plottree=2 plot edges and pause
%  plotgraph=0 no plotting of initial graph (default)
%  plotgraph=1 plot initial graph, without point identification
%  plotgraph=2 plot initial graph, and identify points by indices
% Ids has the indices to be used in identifying the points.
% In each iteration a new class is created, by merging of two classes.
%  Initially each point belongs to a class by itself.
% Each row k of history shows the class that each point belongs to
%  *before* (BEWARE) iteration k.
% Each row k of t shows the action that was taken in iteration k.
%    t(k,:)=[maxall,irow,icol,markrow,markcol,countrow,countcol]
%  where maxall is the correlation of the edge considered;
%  irow,icol are the indices of the 2 points with max correlation
%  markrow, markcol the corresponding classes to be merged
%  countrow,countcol the number of elements in each of the 2 classes
% [t, history]=clustering(a, c, plottree, plotgraph, ids)
if nargin<3, plottree=0; end
if nargin<4, plotgraph=0; end
n=length(a); if nargin<5, ids=1:n'; end
if plotgraph>0, plot(c(:,1),c(:,2), 'ob'); end
if plotgraph>1, identify_points(c(:,1),c(:,2),ids); end
t=zeros(n-1,7);
for k1=1:n, a(k1,k1)=0; end
newmark=n+1; markarray=1:n; k=1; markcount=ones(1,2*n-1);
history=zeros(n,n);
while k<n,
  [maxrow,irow]=max(a);
  [maxall,icol]=max(maxrow);
  irow=irow(icol);
  markrow=markarray(irow); markcol=markarray(icol);
  if markrow~=markcol,
    history(k,:)=markarray;  %%% this is actually the previous step!!
    teamrow=find(markarray==markrow); teamcol=find(markarray==markcol);
    countrow=markcount(markrow); countcol=markcount(markcol);
    tempz=zeros(countrow,countcol);
    a(teamrow,teamcol)=tempz; a(teamcol,teamrow)=tempz';
    markarray(teamrow)=newmark*ones(1,countrow);
    markarray(teamcol)=newmark*ones(1,countcol);
    markcount(newmark)=countrow+countcol;
    t(k,:)=[maxall,irow,icol,markrow,markcol,countrow,countcol];
    if plottree>0,
      hold on; plot(c([irow,icol],1), c([irow,icol],2));
      if plottree>1, pause; end
    end
    newmark=newmark+1; k=k+1;
  end
end
history(k,:)=newmark*ones(1,n); % the last history row
hold off;
```

function v=remove_elements(vall,ind)

```
% returns a shortened version of vall, in which the elements with
indices ind
%   have been removed.
% remove_elements(vall,ind)
vone=ones(1,length(vall));
vone(ind)=zeros(1,length(ind));
v=vall(vone);
```

SCRIPT select20

```
% Used for selecting the corner with the best correlation at time 20,
for each polygon
time_scales_ak=[204*ones(64,1);104*ones(134-64,1); 204*ones(269-134,1)];
gdk=find(time_scales_ak==204)';
maxtime=max(sizes_all_ak(:,2));
vik=zeros(ksofar,1),
vmaxk=zeros(ksofar,1);
hold off;
for k=1:ksofar,
  select_time=round(time_scales_ak(k)*20/204);
  [maxv,maxvirel]=max(vpos(select_time,xlima(k,1):xlima(k,2)));
  maxvi=maxvirel+xlima(k,1)-1;
  vik(k)=maxvi;
  vmaxk(k)=maxv;
end
exselk=extrema(:,vik'); vselk=vpos(:,vik'); nameselk=exnames(vik,:);
'matrix operations'
vposangles=vselk'*vselk;
exangles=exselk'*exselk;
exdistances=2*sin(acos(exangles)/2);
vdistances=2*sin(acos(vposangles)/2);
vdistmax=vmaxk*vmaxk';
chdir Twinky:MATLAB:MikeCBMS:CBMScompressed
save selectv vselk vmaxk exselk nameselk gdk ksofar sizes_all_ak ...
    time_scales_ak xlima exdistances vdistances vdistmax exangles ...
    vposangles vdistmax
```

function [vp,mp,kp]=vprof(kvec)

```
for ik=1:length(kvec),
  k=kvec(ik);
  kk=1;
  while xlima(kk,2)<k, kk=kk+1; end
  %%%eval('global sizes_all_ak vpos xlima');
  kp(ik)=kk;
  mp(ik)=sizes_all_ak(kk,2);
  vp(:,ik)=vpos(1:mp(ik),k);
end
```